Structuration multimodale des vidéos de sport par modèles stochastiques

Ewa Kijak

N° d'ordre: 2941

THÈSE

Présentée devant

devant l'université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention TRAITEMENT DU SIGNAL ET TÉLÉCOMMUNICATIONS

par

Ewa KIJAK

Équipe d'accueil : THOMSON multimedia R&D/Texmex-IRISA
École doctorale : MATISSE
Composante universitaire : IFSIC

Titre de la thèse :

Structuration multimodale des vidéos de sport par modèles stochastiques

soutenue le 22 décembre 2003 devant la commission d'examen

M. :	Claude	LABIT	Président
MM. :	Régine	ANDRÉ-OBRECHT	Rapporteurs
	Bernard	MERIALDO	
MM. :	Lionel	OISEL	Examinateurs
	Philippe	JOLY	
	Patrick	GROS	

Remerciements

Je remercie chaleureusement tous les membres de mon jury : Claude LABIT, directeur de l'IRISA, qui m'a fait l'honneur de présider ce jury ; Régine ANDRE-OBRECHT, professeur à l'Institut de Recherche en Informatique de Toulouse (IRIT) et Bernard Mérialdo, professeur à l'Institut Eurecom, qui ont très aimablement accepté la charge de rapporteurs ; Philippe JOLY, Maître de Conférence à l'Institut de Recherche en Informatique de Toulouse (IRIT) qui a accepté de juger ce travail.

J'aimerais remercier particulièrement Patrick GROS, chargé de recherche au CNRS et responsable du projet TEXMEX de l'IRISA, pour m'avoir permis de réaliser ces travaux de thèse et pour les avoir dirigés. Je tiens à lui témoigner toute ma gratitude pour ses remarques constructives et ses commentaires pertinents. Merci à Gérard BRIAND, responsable de l'ancien Laboratoire AVP de Thomson Multimedia, et à Nour-Eddine TAZINE, qui a pris la relève au sein du Laboratoire CSA, le premier pour avoir initié cette thèse et m'avoir offert l'opportunité de réaliser ces travaux à Thomson Multimedia, le second pour sa confiance et son enthousiasme.

Je tiens tout particulièrement à remercier Lionel OISEL pour son encadrement, ses encouragements et son soutien tout au long de cette thèse. J'ai beaucoup apprécié ses qualités humaines, sa confiance et sa bonne humeur, qui me furent précieuses tout au long de ces travaux. Je remercie également vivement Guillaume GRAVIER avec qui la collaboration entreprise fut pour moi très motivante et très enrichissante. Je lui sais gré de l'interêt qu'il a manifesté pour ce travail, de son implication, de m'avoir accueilli à plusieurs reprises au sein de l'équipe METISS et enfin, d'avoir accepté de faire partie du jury en tant que membre invité.

Je remercie l'ensemble des membres du Laboratoire CSA de Thomson multimédia : François, qui a toujours pris le temps de répondre à mes nombreuses questions avec patience, gentillesse et compétence, Bertrand, Jürgen, Philippe, Louis, Jean-Ronan et Laurence pour m'avoir supporté, pour certains trois années durant.

Je salue également tous les doctorants que j'ai pu rencontrer durant cette thèse, aussi bien ceux de Thomson, de France Telecom, que ceux de l'IRISA, pour des échanges salutaires et enrichissants.

Il me tient enfin à coeur de remercier chaleureusement Teddy, Arnaud et Ronan, amis précieux sans qui ce travail n'aurait pu aboutir à ce qu'il est, pour leurs conseils, leurs encouragements, leur soutien quotidien et jusqu'au boutiste, tout à la fois dominical et véspéral lors de la rédaction du manuscrit.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Ce document est l'issue du travail réalisé dans le cadre d'une convention CIFRE entre le laboratoire Content Search and Access de THOMSON Rennes et le projet TEXMEX de l'IRISA. Certaines parties sont le fruit d'une collaboration avec le projet METISS de l'IRISA.

Indexation vidéo par le contenu

Le contexte général de ce travail est celui de l'**indexation vidéo par le contenu**. L'indexation vidéo est un domaine de recherche né de l'augmentation continue du volume des données multimédia numériques. Dans le domaine audiovisuel, l'utilisation des bases de données multimédia était initialement le fait d'applications professionnelles : archives audiovisuelles, météorologiques, télésurveillance, imagerie médicale, etc. Aujourd'hui, elle concerne également le grand public de par le développement des moyens de mise à disposition de ces contenus : internet, disques durs intégrés aux décodeurs numériques, etc. Nous nous plaçons ici plus particulièrement dans le cadre des vidéos professionnelles télédiffusées (films, émissions), laissant de côté les problématiques liées à la télésurveillance et aux vidéos domestiques ou médicales.

Il est nécessaire de définir des moyens d'accès aux documents qui permettent une utilisation efficace du point de vue des utilisateurs. Différentes modalités de consultation sont considérées : soit en terme de recherche dans une base relativement à une requête formulée par l'utilisateur, soit en terme de navigation dans une base et d'accès rapide à l'information. Ces fonctionnalités peuvent aussi bien être intégrées dans des systèmes de gestion de bases de données professionnelles, que dans des produits à usage domestique comme les magnétoscopes numériques [9].

Ces modalités de consultation s'appuient sur une représentation du contenu. Par analogie avec les documents textuels, différentes formes de représentation du contenu sont envisagées pour un accès rapide à l'information :

une table des index, qui rassemble les mots-clés pertinents correspondant à des requêtes potentielles. À chaque mot-clé est associé un lien vers le document lui-même ;

une table des matières, qui est une structuration hiérarchique du document en parties et sous-parties homogènes en terme de contenu sémantique. Elle contient l'organisation générale et détaillée du document ;

un résumé, qui est une représentation condensée permettant une visualisation rapide de tout ou partie du contenu.

Jusqu'à présent, l'extraction des informations est souvent effectuée par des documentalistes qui associent manuellement à chaque vidéo un certain nombre de mots-clés, qui constituent les index. Il s'agit bien sûr d'une tâche spécialisée très fastidieuse et coûteuse en temps. Ceci montre le besoin d'une extraction automatique des informations. L'indexation vidéo par le contenu est définie comme le processus d'analyse automatique, ou semi-automatique, qui assigne des index décrivant le contenu d'un document vidéo.

Les méthodes d'analyse automatique de la vidéo permettent d'extraire un certain nombre de descripteurs du contenu. Une vidéo peut être analysée à différents degrés de granularité, tels que représentés dans la figure 0.1.

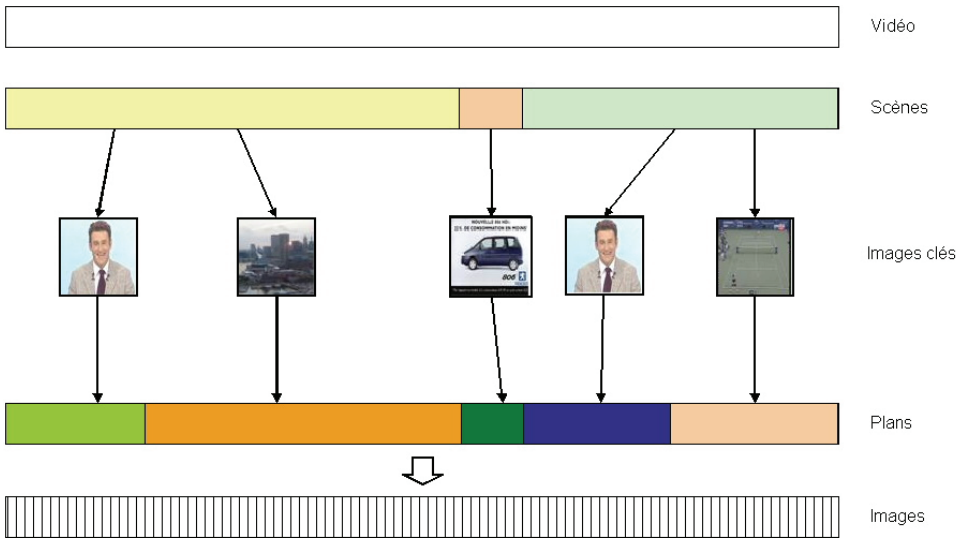


FIG. 0.1: Les différents niveaux de granularité d'un document vidéo.

Le niveau le plus élémentaire est évidemment l'image. Elles sont utilisées pour l'extraction de caractéristiques de couleur, texture et contours.

Au niveau supérieur, les images sont regroupées en *plan*, au sens cinématographique du terme. Un plan est défini par une série d'images acquises de manière continue par une seule caméra. Il représente une action continue dans le temps et dans l'espace. Les plans sont des unités physiques de la vidéo. Ils sont séparés par un effet de montage appelé transition. Les transitions brusques raccordent simplement deux plans successifs. Les transitions progressives regroupent les fondus enchaînés, les volets, les balayages, etc. Cette dernière famille de transition est très variée et leur localisation a fait l'objet de multiples travaux [10, 11, 12, 13]. Chaque plan peut alors être caractérisé par l'extraction d'images-clés représentatives de son contenu visuel, par l'extraction d'entités d'intérêt comme des visages, des zones de texte, ou des objets en mouvement, et par l'extraction d'informations dynamiques comme le mouvement de la caméra.

Au plus haut niveau, les *scènes* regroupent les plans formant une unité narrative (ou unité logique). Le regroupement des plans en scènes est également connu sous le terme de macro-segmentation. Son but est de déterminer une table des matières décrivant la structure du document traité [12]. Appliquée aux films et aux documentaires, l'identification de la structure des documents est également connue sous le nom de *"Inverse Hollywood*

Problem" [14]. Il s'agit de retrouver le script du document à partir de l'analyse de la vidéo. Il est possible de se baser sur les règles de production des films [15], ou des modèles de programmes [16]. Pour ces applications, les plans élémentaires sont des unités trop atomiques et trop peu porteuses de sens. La macro-segmentation est une étape critique de ces systèmes d'analyse. En effet, les scènes sont définies comme un groupe de plans cohérents qui présente un sens pour l'utilisateur. Le problème principal repose sur la définition de la cohérence des plans dans une scène. Est-ce une même unité de lieu, de personne, de thématique ? Contrairement aux plans, la définition des scènes repose sur une corrélation sémantique subjective.

Le problème fondamental de l'analyse de la vidéo par le contenu est le suivant : comment passer de l'extraction d'attributs à un niveau symbolique ? Les approches génériques ne permettent de fournir ni une définition satisfaisante d'une scène, ni une compréhension de haut-niveau des descripteurs extraits.

La seule solution pour extraire des informations de haut-niveau est de restreindre le cadre d'application à une certaine catégorie de documents et d'utiliser des informations *a priori*. Ce sont les *systèmes spécifiques*. En contrepartie d'une compréhension de plus haut niveau, les systèmes spécifiques ont inévitablement un certain nombre d'inconvénients. Ceux-ci sont liés au manque intrinsèque de généralité dû à l'utilisation de règles, de modèles ou d'algorithmes d'apprentissage basés sur une analyse formelle des documents.

Problématique

Parmi les documents télédiffusés, deux catégories ont essentiellement fait l'objet de développements de systèmes spécifiques : les journaux télévisés et les retransmissions d'événements sportifs. Ces documents ont un contenu et une structure bien déterminés. Pour les journaux télévisés, les systèmes spécifiques se sont intéressés à l'analyse de la structure par extraction des unités logiques. Pour les événements sportifs, ils traitent les problèmes de la détection d'événements prédéterminés par le contexte et de la création de résumés.

Dans ce document, nous nous intéressons au problème de l'analyse de la structure des événements sportifs. L'analyse de la structure d'une vidéo sous-entend que le contenu de la vidéo est intrinsèquement structurable, ce qui est le cas des événements sportifs. Cette structure n'est cependant pas directement accessible dans un document vidéo. Connaissant cette structure théorique, le processus de **structuration** consiste à l'extraire automatiquement du document via une série de traitements.

L'identification de la structure est d'une importance capitale pour le processus d'indexation. Elle permet une navigation rapide dans le contenu, grâce à la construction d'une table des matières. Un autre intérêt de la structuration est de réaliser une présélection des segments de la vidéo. Les segments retenus sont ensuite l'objet d'une analyse plus fine comme, par exemple, la détection d'événements particuliers.

L'étude de la structuration des événements sportifs est extrêmement récente et les travaux réalisés dans ce domaine sont pour la plupart contemporains du travail présenté dans ce document.

Si nous nous intéressons à un système aussi spécifique, c'est pour essayer de déterminer le plus haut niveau sémantique qu'il est envisageable d'atteindre en exploitant l'*information*

a priori. Il faut non seulement extraire la structure du contenu, mais aussi identifier sémantiquement chaque composant de cette structure.

Pour atteindre cet objectif, nous devons résoudre trois problèmes principaux. Le premier est d'extraire des informations pertinentes du document audiovisuel. Ce dernier est composé de plusieurs modalités : l'image, le son, et parfois le texte (télétexte, *closed caption stream*). Chacun de ces trois média est riche en informations, et on peut légitimement attendre de leur combinaison qu'elle améliore les performances. La majorité des approches n'utilise cependant qu'un seul média. En effet, l'intégration de plusieurs modalités n'est pas un problème trivial de par leur nature différente. La deuxième difficulté est liée à l'intégration et la représentation dans le système de l'information *a priori*. Enfin, le dernier problème est d'identifier et de représenter la structure de la vidéo, en vue de son extraction.

À une approche heuristique, nous avons préféré un point du vue probabiliste. Notre méthode repose sur un modèle statistique de l'entrelacement temporel des plans de la vidéo. Le cadre général est celui des modèles de Markov cachés. Ceux-ci ont en effet trois avantages : (i) ils proposent un cadre commun propice à l'intégration d'informations multimodales (audio et vidéo), (ii) ils intègrent les informations *a priori*, et enfin (iii) ils permettent de représenter la structure de la vidéo.

Ce document décrit les différents aspects de nos travaux concernant l'analyse de la structure des vidéos de sport. Nous présentons l'état actuel de la recherche sur le développement de systèmes spécifiques dans le chapitre 1. La définition et l'exploitation des informations *a priori* y sont abordées. Cet état de l'art situe également notre travail et notre point de vue dans le contexte plus restreint des systèmes spécifiques.

Le cœur du document est ensuite organisé en quatre chapitres. Le chapitre 2 présente notre méthode de **segmentation d'une vidéo de sport en unités logiques**. Nous définissons les unités logiques, leur **représentation par modèles de Markov cachés** et les descripteurs visuels. La multimodalité est traitée dans le chapitre 3 en intégrant **des indices audio** dans le modèle. L'identification de la **structure hiérarchique** du document, plus complexe et de plus haut niveau que les unités logiques est résolue grâce aux **modèles de Markov cachés hiérarchiques** dans le chapitre 4. Nous verrons que la prise en compte d'une telle structure nécessite l'introduction de descripteurs spécifiques supplémentaires. Nous proposons enfin une nouvelle **perspective** de modélisation de structure de document via les **modèles de segments**.

Chapitre 1

État de l'art sur les systèmes spécifiques

Ce chapitre propose un panorama des systèmes spécifiques dédiés plus particulièrement aux événements sportifs. Après une description générale du contexte de l'analyse des vidéos de sports, nous décrirons quelles sont les informations *a priori* utilisées par les systèmes spécifiques. Nous expliquerons comment elles sont exploitées à la fois dans l'analyse bas-niveau et haut-niveau des vidéos de sports, et quels types de connaissance elles permettent d'extraire. Nous établirons un parallèle entre les informations *a priori* et les méthodes utilisées pour l'analyse des événements sportifs et celles utilisées pour l'analyse des journaux télévisés. Enfin, l'ensemble des techniques évoquées sera discuté afin de situer le point de vue que nous avons privilégié pour notre étude.

1.1 Introduction

Le problème fondamental lié aux systèmes d'indexation est le manque de coïncidence entre la simplicité des informations que l'on peut extraire automatiquement de données visuelles et l'interprétation complexe de ces mêmes données par un utilisateur dans une situation donnée. Ce problème est plus connu dans la littérature sous le nom de "semantic gap". Pour tenter de combler ce fossé, la plupart des approches s'intéressent à l'intégration et à l'évaluation d'informations spécifiques à un domaine donné.

Il existe différents niveaux de granularité des index que l'on classe en quatre niveaux hiérarchiques [17] : le genre, le sous-genre, les unités logiques (ou scènes) et les événements. Les systèmes spécifiques se placent dans un genre identifié, par exemple celui des événements sportifs ou des journaux télévisés. Les objectifs sont alors les suivants :

- classer en sous-genre, c'est à dire identifier le sport diffusé (tennis, volley, football, baseball, etc.) ;
- pour un sport donné (sous-catégorie fixée), extraire les scènes (par exemple, phases de jeu ou de non jeu). Pour les journaux télévisés, distinguer les scènes de plateau des reportages ;
- pour un sport donné, détecter des événements particuliers (buts pour le football, paniers pour le basket, etc).

L'étude des systèmes spécifiques est un domaine de recherche récent. Les premiers

articles de référence présentant des systèmes dédiés aux journaux télévisés [16, 18] et aux événements sportifs [19, 20, 21] datent d'une dizaine d'années. La recherche s'intensifie dans les années 2000. Les promesses offertes par l'exploitation de la connaissance *a priori* pour atteindre un niveau de compréhension sémantique de la vidéo, la composition bien déterminée du contenu de ces vidéos, et enfin l'enjeu applicatif que représentent ces deux domaines à eux seuls - dû à une importante production - suscitent l'intérêt des centres de recherche, aussi bien que celui des entreprises. Le nombre croissant de publications a donné lieu à la première session spéciale dédiée à l'analyse des vidéos de sports dans une conférence internationale ICIP (International Conference on Image Processing) en septembre 2003, prouvant l'intérêt croissant de la recherche et la reconnaissance à part entière de ce domaine.

Ce chapitre propose donc une vue générale des travaux relatifs aux systèmes spécifiques d'analyse des événements sportifs. La plupart des travaux cités ont été publiés ces trois dernières années.

1.2 Exploitation des informations *a priori*

L'intérêt de l'analyse des vidéos de sports télédiffusées a déjà été évoqué ci-dessus :
- importance de la production, donc des données concernées ;
- événements possibles bien identifiés ;
- nombre limité de ces événements.

L'énorme production des vidéos de sport est à l'origine de l'importance des applications liées à leur analyse. Le nombre limité d'événements bien identifiés implique que le contenu soit compressible dans le temps, autrement dit, qu'il pourrait se résumer à l'ensemble de ces événements. Globalement, l'objectif final est d'être capable de proposer à un utilisateur soit un résumé des meilleurs moments d'une rencontre sportive, soit un index lui permettant de sélectionner les segments qu'il veut visualiser.

L'objectif commun de ces systèmes d'analyse est d'identifier tout ou partie des segments intéressants d'une vidéo. La nature de ces segments (événements) diffère selon le type de sport traité. Par exemple, pour le football, le baseball, le basketball, on essayera de retrouver les séquences correspondants à des points marqués. Pour le tennis, le billard, le patinage artistique, on identifiera les différentes figures.

Pour atteindre ces objectifs, les systèmes ont besoin de s'appuyer sur des informations *a priori*. Nous distinguons deux types d'informations *a priori* pouvant être exploitées :

1. les informations liées à la nature du sport considéré ;
2. les informations liées à la production des vidéos télédiffusées.

Dans cette partie, nous détaillons la nature de ces informations, les méthodes utilisées pour les obtenir et la connaissance qu'elles permettent d'extraire.

1.2.1 Informations *a priori* liées au domaine

Les informations liées au domaine sont les connaissances intrinsèques à la nature et aux règles du sport étudié, telles que :
- la surface de jeu : couleur et composition du terrain, géométrie et couleurs des lignes ;
- le nombre de joueurs ;

– le déroulement du jeu.

Connaissant le modèle du terrain, la détection et la reconnaissance du marquage au sol permettent de localiser le jeu sur le terrain [20], en particulier lorsque celui-ci est trop vaste pour entrer dans le champ de la caméra, comme pour le football. Les joueurs, la balle ou le ballon peuvent être segmentés et suivis.

1.2.1.1 Identification de la couleur du terrain

Bien que la couleur du terrain ne soit pas invariante avec les conditions d'illumination et la profondeur de champ de la caméra, une connaissance initiale de cette couleur aide à identifier les régions de l'aire de jeu par une segmentation basée couleur. Quelque soit le sport, le modèle de la couleur du terrain est déterminé à partir de l'analyse des plans larges pour lesquels le terrain est caractérisé par une région uniforme dominante dans l'image. À partir des images des plans larges, les caractéristiques couleurs du terrain sont donc déterminées par des techniques de clustering. Cela suppose d'avoir identifié les plans larges parmi l'ensemble des plans, et cela sans connaissance *a priori* de la couleur du terrain.

Pour résoudre ce problème, la couleur du terrain est fixée *a priori* en fournissant une image modèle d'un plan large, une région de l'espace couleur [22], ou un histogramme couleur du terrain [23, 24]. Une autre méthode consiste à construire un ensemble de modèles de couleurs de terrain possibles par apprentissage. La couleur du terrain de la vidéo étudiée est ensuite identifiée par une mesure de distance à chaque modèle [25, 26].

Une autre façon de résoudre ce problème est d'identifier les plans larges en utilisant par exemple une estimation de mouvement. L'activité locale et le mouvement global persistant de la caméra permettent de filtrer les gros plans et les suivis [27]. La détection des lignes du terrain permet également de caractériser les plans larges. Li *et al.* [28] recherche des images de couleur dominante proche du vert possédant des lignes blanches parallèles pour identifier les plans larges d'une vidéo de football américain et estimer ainsi la couleur réelle du terrain. Cette approche basée image suppose néanmoins de détecter et analyser les lignes de toutes les images candidates, ce qui se révèle fastidieux. La moyenne sur le plan du moment d'ordre 2 de la transformée de Hough a également été utilisée pour identifier respectivement les vues d'une table de snooker ou les vues d'un terrain de tennis [29, 30]. Lorsque les plans larges sont identifiés, la couleur du terrain est estimée par une phase d'apprentissage sur la vidéo traitée [31, 32, 28, 27]. La couleur du terrain étant soumise à des variations d'illumination, sa couleur est mise à jour de manière adaptative au long de la séquence [32].

1.2.1.2 Identification des lignes du terrain

L'identification des lignes du terrain exploite intensivement la géométrie connue du terrain, la couleur des lignes (en général blanches), ainsi que leur longueur et leur orientation [28].

Pour la détection des lignes, une approche classique consiste à effectuer un filtrage sur la couleur des pixels pour séparer le terrain, les lignes du terrain, et les joueurs, couplée à une détection de contours. La détection des lignes est alors réalisée en utilisant la transformée de Hough [23, 28, 32, 25], ou un chaînage des contours exploitant la couleur blanche des lignes [20]. Des algorithmes spécifiques de suivi de droites exploitant la connaissance *a priori* de la direction de recherche sont également développés [26].

Les lignes sont alors identifiées grâce à la connaissance du modèle du terrain. Pour les sports dont le terrain est trop étendu pour être capturé entièrement par une caméra, l'identification des lignes localise le jeu selon des classes prédéfinies. Les exemples suivants sont relatifs à l'analyse des vidéos de football. Les classes sont caractérisées simplement par le motif du marquage au sol et identifiées par le calcul d'une signature caractérisant ce motif [20]. L'absence de marquage au sol dans certains plans de jeu est compensée par une analyse du mouvement. Cependant, sans identification des différents angles de la caméra et du facteur de zoom, la reconnaissance du marquage n'est pas robuste. Les différentes zones de jeu sont également caractérisées par leur prise de vue : forme de l'aire de jeu, orientation des lignes du terrain, pourcentage de pixel de la couleur du terrain, présence d'un angle et d'une ligne centrale, identifiées avec un classifieur bayésien naïf [1].

Sans vouloir identifier à chaque instant la localisation du jeu sur le terrain, la recherche et la reconnaissance de certains motifs particuliers du marquage au sol peut être utile. Par exemple, la présence de trois lignes parallèles permet de détecter la zone de but [33, 32].

Une fois les lignes identifiées, il devient possible de calculer la déformation entre l'image et un modèle théorique du terrain. Cette déformation permet alors de localiser dans un réferentiel d'autres objets détectés. Pour déterminer la position du ballon et des joueurs sur le terrain, Choi *et al.* [23] détectent d'abord le cercle central du terrain pour calculer la transformation homographique entre les points extraits de l'image et le modèle du terrain. Lorsque le cercle central n'est pas présent dans l'image, une image mosaïque est utilisée pour étendre le champ de recherche. Dans ce cas, la transformation en mosaïque est combinée avec la transformation homographique, ce qui alourdit l'approche.

1.2.1.3 Détection et suivi des objets

Dans le cas des sports d'équipe, la détection et le suivi des joueurs est une tâche délicate à cause du nombre variable de joueurs présents dans le champ de la caméra, de leur position variable, des occultations et de la mauvaise qualité de la vidéo.

Un filtrage sur la couleur du terrain est souvent utilisé pour extraire des blobs, définis comme les blocs de pixels dans la surface du terrain n'ayant pas la couleur du terrain. Les joueurs sont alors identifiés et différenciés par la couleur de leur maillot. De nombreuses approches utilisent un histogramme de la couleur des maillots pour modéliser les joueurs de chaque équipe. L'affiliation des joueurs est alors déterminée par une mesure de similarité [23, 34, 28, 35]. Certains travaux identifient même l'arbitre, dont la présence en gros plan est caractéristique de certains événements [32].

Le suivi des joueurs est généralement basé sur des techniques de "template matching" [20, 21, 23, 34, 36, 33, 26] qui ne gèrent pas en général les occultations lorsqu'il s'agit de deux joueurs de la même équipe (donc pour lesquels une distinction sur la couleur du maillot est impossible). La prédiction du mouvement des joueurs par extrapolation linéaire en utilisant leur vitesse moyenne permet de mieux gérer les occultations [22]. Lefevre [11] a proposé une méthode rapide de suivi par contours actifs adaptée aux petits objets, admettant une phase de scission utile dans le cas d'objets proches et pouvant s'occulter.

La difficulté du suivi dans le cadre d'un sport d'équipe nécessite d'intégrer des connaissances *a priori* pour résoudre les incertitudes inhérentes au système de suivi. Pour le suivi des joueurs de football américain, Intille *et al.* [19] proposent une méthode, basée sur l'hy-

pothèse de "monde fermé", qui intègre des informations contextuelles telles que le nombre et le type d'objets sur le terrain.

Le suivi de la balle/du ballon est également un problème difficile car il s'agit souvent d'un objet de petite taille qui bouge très vite et dont les occultations sont fréquentes. LucentVision [37] et ESPN K-Zone [38] se sont intéressés au suivi d'objets spécifiques pour le tennis et le baseball, respectivement. Le premier analyse la trajectoire des deux joueurs et de la balle, et le second suit la balle durant les lancers pour montrer si les décisions de l'arbitre concernant le lancer sont correctes. Le temps réel dans ces deux systèmes est atteint en utilisant intensivement la connaissance *a priori* sur le système de capture mis en place telle que la localisation des caméras, leur calibration et le champ qu'elles couvrent. Ces systèmes fonctionnent en amont de la production, tandis que l'analyse des vidéos télédiffusées se situe en aval.

1.2.2 Informations *a priori* liées à la télédiffusion

Les informations liées à l'édition de la vidéo pour la télédiffusion sont largement utilisées pour l'analyse de la vidéo. L'édition d'une vidéo correspond au choix des plans et à leur montage avant diffusion. Cette étape effectuée par un opérateur humain, loin d'être innocente et aléatoire, est lourde de sens.

En effet, les événements sportifs sont filmés à partir d'un nombre fixe de caméras réparties autour de l'aire de jeu afin de pouvoir capturer au mieux l'action. Pour la diffusion, c'est bien entendu le point de vue (la caméra) rendant le mieux compte de l'action en cours qui est sélectionné. De cette règle élémentaire à la base du processus d'édition découle un ensemble de règles de production, partagé par toutes les vidéos de sport professionnelles.

1.2.2.1 Prise de vue et classification des plans

Parce qu'il y a un nombre fixe de caméras, il y a un nombre fixe de points de vue recensés. Parce que le meilleur point de vue, au sens de l'action pour le sport considéré, est sélectionné à chaque instant, l'identification des différents points de vue fournit une information sur le statut du jeu. C'est selon ce raisonnement commun qu'une majorité d'approches s'intéressent à la classification des prises de vues. Dans les versions les plus optimistes, cette classification est la finalité du système d'analyse. Plus généralement, elle est considérée comme la première étape incontournable d'un système d'analyse.

Quelque soit le sport considéré, on distingue quatre classes majeures :

plan d'ensemble/général/vue du terrain (ou *large/global/field view*) : il s'agit de plans larges fournissant une vue globale de l'aire de jeu.

plan moyen (ou *medium view*) : il désigne un zoom de la caméra sur un champ restreint de l'aire de jeu, ou un cadrage en pied du sujet.

plan rapproché et gros plan (ou *close-up*) : il désigne un cadrage à hauteur de poitrine ou du visage sur le joueur, le public ou l'arbitre.

vue du public (ou *audience view*)

Suivant les sports considérés, cette classification est plus ou moins détaillée, notamment en distinguant plusieurs types de plans rapprochés. Toujours selon le sport considéré,

l'interprétation des classes, souvent qualifiée de "sémantique", varie. Le tableau 1.1 issu de [27] résume pour quatre sports (tennis, football, basketball et volleyball), l'interprétation sémantique des différentes prises de vue, que l'on retrouve dans l'ensemble de la littérature.

Pour chaque sport, un ou plusieurs types de prise de vue sont caractéristiques des phases de jeu et de non-jeu. Ces vues sont aussi respectivement désignées dans la littérature par les termes "vues actives" et "vues non-actives". Par exemple, au baseball, les vues du lanceur sont typiquement le point de départ d'un événement. Au tennis, ce sont les vues du terrain qui contiennent l'engagement d'un nouveau point.

Le point de vue des plans véhicule une certaine information sur le statut du jeu. La majorité des approches basées plan les classifient selon le point du vue représenté. L'identification des plans est également utilisée comme pré-sélection à une analyse plus fine de leur contenu, réalisée par la détection du terrain et le suivi des joueurs. Par exemple, Gong *et al.* [20] proposent, pour les vidéos de football, une reconnaissance des motifs des marquages au sol du terrain pour classer la position du jeu. Ils ne disposent d'aucun système sélectionnant les segments à traiter, entre les vues du terrain et les gros plans. Sans identification préalable des segments, l'algorithme doit gérer beaucoup plus de bruit. De la même façon, Miyamori [39] ne traite que les segments contenant une vue globale du terrain de tennis, afin d'analyser la position des joueurs, de la balle et du terrain pour identifier des coups du tennis. L'identification des vues du terrain est supposée avoir été préalablement réalisée.

La classification des plans s'appuie sur l'analyse des attributs bas-niveau extraits de la vidéo, principalement la couleur, mais aussi le mouvement de la caméra et les contours. Un certain nombre d'approches combinent plusieurs caractéristiques pour classifier les plans. Cette classification est réalisée :

- selon des règles heuristiques [40, 41] : une région bleue au milieu de la partie gauche, et des contours horizontaux et verticaux autour de cette région indiquent la partie gauche du terrain de basketball. Une région dominante orange et au moins deux lignes verticales et horizontales indiquent une vue du terrain de tennis.
- par réseau de neurones [42, 43]
- par arbre de décision [27, 43]
- par classifieur Bayésien [44] : chaque classe de plans est représentéé par les densités de probabilités des différents attributs.

Utilisation de la couleur

Une fois la couleur du terrain identifiée, une méthode simple de classification consiste à prendre une décision sur le taux de pixels par image ayant la couleur du terrain [31, 33]. Ce critère heuristique simple ne distingue cependant pas toujours précisément les plans larges des plans moyens. Pour ôter les incertitudes, il est combiné à une détection des blobs [33] ou à la "Golden Section Rule" qui est une règle de cadrage de l'image en fonction des sujets [32].

Une autre méthode consiste à prendre une décision sur une mesure de similarité entre l'histogramme couleur des images-clés et l'histogramme modèle du terrain, lorsqu'il a été estimé ou fourni. Zhong *et al.* [25] sélectionnent des images-clés candidates sur la similarité de leur histogramme couleur au modèle de la vue active du baseball ou du tennis. Ils

Type	Classe	Taux présence	Classes prépondérantes	Interprétation sémantique
Tennis (20 mn)	Gros plan	37.5	*	Temps mort
	Vue du court	29.8	*	Jeu
	Plan moyen d'un joueur	16.8	*	Temps mort
	Public	9.2		Réaction active
	Vue hélicoptère	1		Début d'un nouveau jeu
	Rediffusion	4.3		Événement
	Indéfini	1.4		
Football (45 mn)	Gros plan	26	*	Temps mort
	Vue du terrain	33.6	*	Jeu
	Suivi	22.1	*	Jeu
	Plan moyen statique d'un joueur	3		Coup-franc, Penalty
	Public	0.7		Réaction active
	Corner	0.7		
	Vue des buts	4.8		Attaque
	Rediffusion	4.1		Evénement
	Indéfini	5		
Basketball (18 mn)	Gros plan	41.9	*	Temps mort, lancer franc
	Full Court Advance	22.6	*	Jeu
	Vue statique du terrain	5.4		Début d'un nouveau jeu
	Plan moyen d'un joueur	4.3		Événement
	Public	7.5		Réaction active
	Penalty	7.5		
	Vue hélicoptère	2.2		Début ou fin d'un jeu
	Rediffusion	7.5		Evénement
	Indéfini	1.1		
Volleyball (22 mn)	Gros plan	48	*	Temps mort, service
	Vue du terrain	36.5	*	Jeu
	Vue statique du terrain	3.8		Service
	Public	1.9		Réaction active
	Rediffusion	5.8		
	Indéfini	3.8		

TAB. 1.1: Signification des prises de vue résultant de l'analyse de différentes vidéos de sport.

procèdent ensuite à une étape de vérification basée sur la segmentation des objets, la détection des contours et leur adéquation avec des connaissances *a priori* telles que la présence d'un objet en mouvement de la "bonne" taille au "bon" endroit et la présence d'au moins deux lignes horizontales et verticales séparées d'une certaine distance.

Une autre hypothèse exploitée est que les deux classes majoritaires de points de vue dans une vidéo sont les gros plans et les plans larges [45]. Ces classes sont identifiées à partir de la constuction d'une matrice de similarité de tous les plans, la similarité $d(S_i, S_j)$ entre deux plans S_i et S_j étant définie comme l'intersection des histogrammes des couleurs dominantes des images-clés respectives. Pour une vidéo contenant N plans, cette approche nécessite le calcul de N^2 mesures de similarité pour construire la matrice, ce qui s'avère coûteux pour des vidéos de taille importante.

Utilisation des contours

L'information de contours est parfois également considérée pour identifier les plans sur la base d'hypothèses comme une forte présence de contours dans les plans sur le public [46]. Cette mesure seule n'est cependant pas très convaincante. Couplée à l'extraction des couleurs dominantes, la distribution des contours selon quelques directions privilégiées classe les plans d'une vidéo de basketball en quatre catégories : vue de la partie gauche du terrain, vue du centre, vue de la partie droite et autres (gros plans) [40].

Utilisation du mouvement de la caméra

Tout comme la couleur du terrain, le mouvement de la caméra est un attribut souvent considéré pour caractériser les plans. Il présente l'avantage par rapport à l'information couleur de ne pas requérir d'informations *a priori* sur la couleur du terrain, d'éviter une segmentation complexe de l'image, et d'être extrait rapidement lorsqu'il est estimé dans le domaine compressé [40, 47, 48, 49].

En effet, au même titre que l'identification des points de vue, l'analyse du mouvement de la caméra contient une forte information sur le statut du jeu. Il existe une corrélation entre le mouvement de la caméra et le mouvement du ballon ou de la balle puisque l'action dépend de la position de ce dernier, et parce que le meilleur point de vue doit être sélectionné à chaque instant. Le mouvement de la caméra suivant celui de la balle, des événements sont directement déduits de l'observation du mouvement global d'une scène.

Les actions caractéristiques d'une caméra sont les mouvements de translation, rotation (*tilt*) et zoom. La translation est utilisée pour déplacer le champ d'un endroit à un autre du terrain, tandis que le zoom change le sujet du point du vue.

L'analyse du mouvement dominant de la caméra durant un plan est utilisée pour localiser l'action sur le terrain pour le cricket [50]. Quant à l'évolution temporelle des paramètres du mouvement, elle permet d'identifier des événements à partir de règles heuristiques. Par exemple pour le cricket, une frappe du batteur sera caractérisée par un zoom significatif suivi d'un changement de direction du mouvement de la caméra. La direction de la frappe est mesurée par la direction de la translation qui suit ce zoom significatif [51]. Ce n'est alors pas la précision de l'estimation de mouvement mais le changement temporel dans l'information de mouvement qui est importante.

De façon générale, les valeurs numériques des paramètres de mouvement sont souvent quantifiées et/ou converties en valeurs symboliques qualifiant le mouvement selon son in-

tensité (faible, moyenne ou forte) et/ou sa direction (droite, gauche, haut, bas) [1, 40].

Ces approches ne prennent pas en compte la notion temporelle de la succession des plans. Or, toutes les classes ne peuvent cependant pas être estimées à partir d'un seul plan (les corners au football [27], par exemple). Le contexte doit alors être exploité.

1.2.2.2 Règles de montage

Les règles de montage découlent de la règle élémentaire qu'est la sélection du meilleur point de vue à chaque instant. Cela implique notamment qu'un gros plan ne sera pas diffusé pendant une action de jeu, et par extension les publicités non plus - avec une nuance sur ce dernier point pour les pays gros diffuseurs de publicités. Cela implique également qu'après une action d'éclat, le point du vue sélectionné sera un gros plan sur les joueurs ayant exécuté l'action, ou une vue du public. En résumé, en plus du type de point de vue, l'analyse de l'entrelacement temporel des plans fournit des indications sur le déroulement du jeu.

Au choix de la succession des plans s'ajoute, lors de la phase d'édition, l'insertion des rediffusions. Les rediffusions proposent des ralentis de la scène précédente, ou plus simplement un autre point de vue, et elles interviennent immédiatement après qu'un événement a été jugé suffisamment intéressant. Elles sont alors précédées ou encadrées de transitions spéciales dont le rôle est de notifier au téléspectateur qu'il s'agit d'une rediffusion. Les effets de transitions sont des volets, des fondus enchaînés, des balayages, ou toute autre forme de transition progressive.

Les moments importants pendant un programme de sport sont souvent rediffusés au ralenti juste après qu'ils ont eu lieu. Plusieurs approches se sont intéressées à la détection des ralentis afin de localiser les moments importants [52, 53, 27]. Ceux-ci sont ensuite concaténés pour construire des résumés [54].

La figure 1.1 représente un diagramme simplifié de la structure d'un segment contenant un ralenti dans une vidéo de sport. Le plan contenant l'action est souvent suivi d'autres plans avant la rediffusion au ralenti, qui contient elle-même des effets de transitions au début et à la fin.

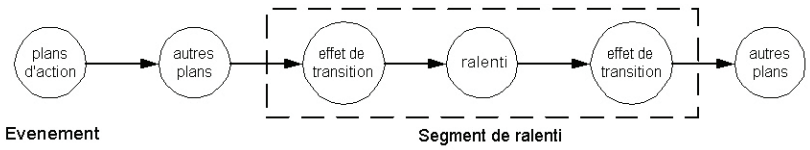


FIG. 1.1: Structure d'un segment contenant un ralenti dans un programme de sport.

Détection des ralentis

Si les ralentis sont enregistrés par des caméras standards, alors certaines images sont simplement répétées pour donner une impression de ralenti.

Cette redondance des images générée par un ralenti est utilisée pour détecter les ralentis [53, 55]. L'absence totale de mouvement entre deux images identiques successives est détectée au moyen des vecteurs mouvement MPEG. La localisation des segments de ralentis n'est cependant pas adressée dans ce cas.

L'utilisation d'un modèle de Markov caché (HMM) permet à la fois de modéliser les relations entre les différents types d'images composant un ralenti et de localiser les frontières [54]. Le HMM représenté à la figure 1.2 est construit de façon à modéliser la moitié d'un ralenti. Les états (0-3) correspondent au type d'images présentes dans un ralenti, tandis que l'état (4) est l'état de sortie qui localise la fin du segment.

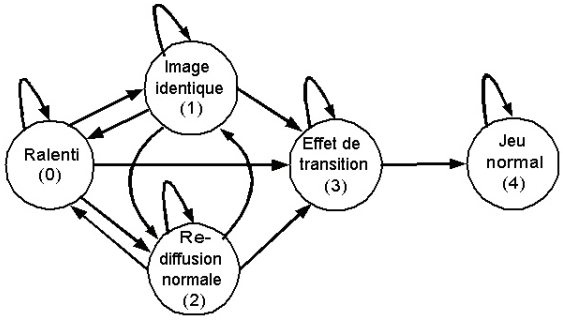


FIG. 1.2: Structure du modèle de Markov représentant un ralenti dans un programme de sport.

Selon cette méthode, la localisation et la détection sont impossibles si une caméra rapide est utilisée, auquel cas les images du ralenti sont diffusées à allure normale, ou si les images sont sous-échantillonnées pendant l'encodage.

Détection des transitions spéciales

Plutôt que de détecter directement les rediffusions (pas seulement les véritables ralentis, mais aussi les rediffusions de l'action sous un autre angle), d'autres approches cherchent les transitions spéciales au début et à la fin des rediffusions.

La détection des transitions spéciales n'est pas triviale. D'abord, la nature généralement progressive de ces transitions ne permet pas d'appliquer de simples techniques de comparaisons successives d'images. Ensuite, la mise en œuvre de ces transitions varie d'une vidéo à l'autre, tantôt fondu enchaîné, tantôt volet, tandis que les algorithmes de détection sont spécifiques à un type de transition.

Aussi les transitions spéciales sont considérées connues *a priori* : balayage dont la manipulation de l'image est connue a priori [52], fondus enchaînés utilisant une image de synthèse (Fig. 1.3) [27] ou un logo [56]. Dans ces deux derniers cas, les modèles graphiques des transitions sont appris à partir de la séquence traitée. Les transitions sont alors identifiées par comparaison des images successives au modèle. La comparaison de toutes les images de la vidéo est coûteuse en temps de calcul [27]. Aussi, Pan *et al.* [56] restreignent l'ensemble de recherche en détectant d'abord les ralentis grâce à la méthode non robuste décrite au paragraphe précédent [54].

FIG. 1.3: Exemple d'une transition spéciale utilisant une image de synthèse.

1.2.2.3 Incrustations

Le dernier élément propre à la télédiffusion est l'incrustation d'informations, en général relatives au score, au nom des joueurs, ou aux statistiques du jeu. Certaines informations sont incrustées de façon permanente, comme le score et le nom des équipes. Les autres informations apparaissent de façon ponctuelle : nom du joueur venant de mener une action, statistiques.

L'avantage de la reconnaissance des incrustations est le fort niveau sémantique que l'on peut immédiatement en tirer. Elle fournit des informations sur l'état du jeu à chaque événement (score, nombre d'essais, ...).

L'agencement spatial des incrustations et l'organisation des informations dans la zone incrustée est propre à chaque sport et à chaque diffuseur de contenu. Il est donc nécessaire d'initialiser le système de détection pour s'adapter à chaque nouveau flux. L'initialisation consiste à trouver la localisation du score dans l'image, et les masques des différentes zones contenant le texte à analyser. Il faut également connaître ou apprendre l'information portée par chaque région de la zone de texte (score, nom de l'équipe, ...), plus ou moins nombreuses selon les sports. Ces régions sont fournies *a priori* [2, 24] ou identifiées automatiquement. Une façon d'identifier ces régions une fois qu'elles sont détectées, est d'utiliser les fréquences de changement du texte et de déduire des règles liées au sport et des règles d'édition : quelle région est dédiée au nom des équipes, quelle autre au score, etc... [57].

Les connaissances *a priori* sur l'évolution du score (par exemple si le score courant est 1-1, le score suivant devrait être 1-2 ou 2-1) améliorent le système de reconnaissance, en corrigeant les fausses détections. Elles peuvent être intégrées sous la forme d'un graphe de transition [58].

La reconnaissance du texte incrusté dans les différentes régions détecte les événements simplement : un changement du score indique nécessairement qu'une action importante vient d'être menée. Une analyse visuelle de la vidéo détermine alors le segment vidéo correspondant à l'événement.

Nous venons de décrire les informations *a priori* utilisées par les systèmes spécifiques et le type d'informations qu'elles permettent d'extraire. Nous allons voir dans la section suivante comment elles sont exploitées et avec quels objectifs.

1.3 Analyse haut-niveau des retransmissions sportives

Comme il l'a été mentionné dans l'introduction, on distingue quatre niveaux dans l'analyse des vidéos :

1. le genre ;
2. le sous-genre ;
3. les unités logiques ou scènes ;
4. les événements.

Nous nous situons dans le cadre d'un genre spécifique, celui des vidéos de sport. A chaque niveau de granularité correspond un objectif différent. Au niveau du genre, l'objectif est d'identifier les vidéos des sports recherchés parmi un ensemble varié de contenus. Au niveau du sous-genre, l'objectif est d'identifier différents types de sports parmi un ensemble de vidéos de sports. Au niveau des unités logiques, l'objectif est d'identifier des segments de la vidéo, *i.e.* un ou plusieurs plans, qui forment une unité narrative ou structurelle, appelées *unités logiques*. Pour les vidéos de sport, les unités logiques sont par exemple les phases de jeu et de non-jeu. Au niveau des événements, l'objectif est d'identifier un événement prédéterminé dans une vidéo donnée. Dans le cadre des événements sportifs, les événements sont prédéterminés par la nature du sport traité :

- pour le football : détection de buts ;
- pour le basketball : détection des paniers :
- pour le baseball : les frappes réussies ;
- pour le tennis : les coups joués (revers, coup droit, smash) ou les différents points marqués (ace, montées au filet).

On entend par "événements" des actions ponctuelles dans la vidéo par opposition aux "unités logiques". Les événements sont à la base de la construction d'une *table des index*, quand les unités logiques sont le fondement de la *table des matières*. Après avoir présenté la classification en sous-genre, nous nous intéresserons plus particulièrement dans cette section à la détection d'événements et à la segmentation en unités logiques. Pour ces deux catégories, nous présenterons les approches déterministes et les approches probabilistes.

1.3.1 Classification des sports

La classification des sports a pour but d'identifier les segments d'une vidéo relatifs à un même sport. Des modèles des classes caractérisant chaque sport sont construits et la probabilité d'un segment d'appartenir à chacune des classes est calculée. Ces modèles sont des densités de probabilité des attributs utilisés par une estimation bayésienne [59, 55], intégrés dans des HMMs [60], ou des machines à états finis [43]. Le segment est classifié selon le meilleur score obtenu.

La classification est appliquée pour distinguer différents sports au sein de magazines de sports contenant différents reportages [42, 61], mais aussi pour distinguer les différentes disciplines d'une vidéo d'athlétisme [43]. Plus généralement, Kobla *et al.* [55] distinguent les segments relatifs aux événements sportifs dans un flux télévisé, sans cependant essayer d'identifier le sport représenté. Les segments de sport sont caractérisés par une forte fréquence d'apparition de texte dans les images, la présence de nombreux ralentis et une quantité de mouvement plus importante que pour les segments non relatifs au sport.

Dans le cas des magazines de sport, les plans consécutifs correspondant à un même sport sont regroupés à partir d'une mesure de corrélation des plans et d'un regroupement contraint temporellement [61]. Une approche alternative consiste à classifier les plans individuellement. Une préclassification supprime les scènes de plateau relatives au présentateur et aux interviews. Les plans restants sont alors classifiés selon le type de vue qu'ils représentent. Étant donné un modèle de couleur du terrain caractérisant chaque sport, la reconnaissance des sports est réalisée par l'analyse de la couleur sur les vues du terrain [42]. Afin d'éviter d'avoir recours à la couleur qui, pour un même sport, peut varier d'une séquence à l'autre, les seuls paramètres de mouvement et les transitions entre les différents mouvements estimés peuvent être utilisés d'après une analyse statistique réalisée sur six sports [62].

1.3.2 Détection d'événements

Les événements sont définis par rapport au contexte *i.e.* au sport considéré. Pour détecter les événements, la dimension temporelle de la vidéo est utilisée. Elle représente l'évolution du jeu à travers le temps et se traduit par la succession des différents plans et l'évolution du mouvement de la caméra, autrement dit par les règles de production.

1.3.2.1 Approches déterministes

Dans les approches déterministes, les règles sont interprétées explicitement en terme de règles de décision.

Ces règles sont regroupées et interprétées dans un module de raisonnement. Le tableau 1.2 représente un module de raisonnement qui analyse la position des deux joueurs par rapport au terrain pour identifier des événements tels que jeu de fond de court, service volée, passing shot dans les vues du terrain d'une vidéo de tennis [26]. Cette classification est cependant ambiguë : un plan dans lequel les joueurs restent près de la ligne de fond de court est-il un échange de fond de court, ou les joueurs se préparent-ils à servir ?

joueur 1		joueur 2		événement
position initiale	position finale	position initiale	position finale	
FC	FC	FC	FC	échange fond de court
FC	F	FC	FC	passing-shot
FC	FC	FC	F	passing-shot
FC	FC	CFC	CLS	service-volée
CFC	CLS	FC	FC	service-volée
LS	F	LS	F	jeu au filet

TAB. 1.2: Module de classification en événements d'une vidéo de tennis. FC = Fond de Court, LS = Ligne de Service, F = Filet, CFC = Centre de la ligne de Fond de Court, CLS = Centre de la Ligne de service.

Un module de raisonnement identifie également les actions des joueurs au tennis parmi revers, coup droit et smash dans [36, 39]. Cette fois, ce sont la position des joueurs et de la balle au moment de l'impact entre la raquette du joueur et la balle qui sont analysées.

Les règles sont simples : si la balle est à gauche, il s'agit d'un revers (ces règles ne tiennent cependant pas compte de la prédisposition du joueur à tenir sa raquette de la main gauche ou de la main droite). Mais le suivi et la détection des joueurs et de la balle image par image sont des traitements complexes entraînant des temps de calcul prohibitifs (plus d'une demi-journée pour une heure de vidéo).

L'analyse de la position du ballon et des joueurs par rapport à la zone de but est également utilisée pour la détection des buts au football [11].

Le recours à un arbre de décision est une façon d'intégrer facilement des règles et des informations *a priori*. Par ailleurs, il montre clairement l'ordre et la priorité de chaque attribut utilisé et assure une classification rapide. Il faut néanmoins décider à chaque noeud quels attributs et quelle règle utiliser. La figure 1.4 montre un exemple d'arbre de classes sémantiques permettant de classifier les plans d'une vidéo de basketball selon neuf événements prédéterminés (les feuilles de l'arbre). La structure de l'arbre est déterminée manuellement de façon à réaliser les décisions les plus discriminantes en premier. Les attributs et le seuil utilisés sont estimés par un apprentissage supervisé à chaque noeud [40]. Cependant, la classification est réalisée ici plan par plan, sans tenir compte d'aucune information sur le voisinage des plans.

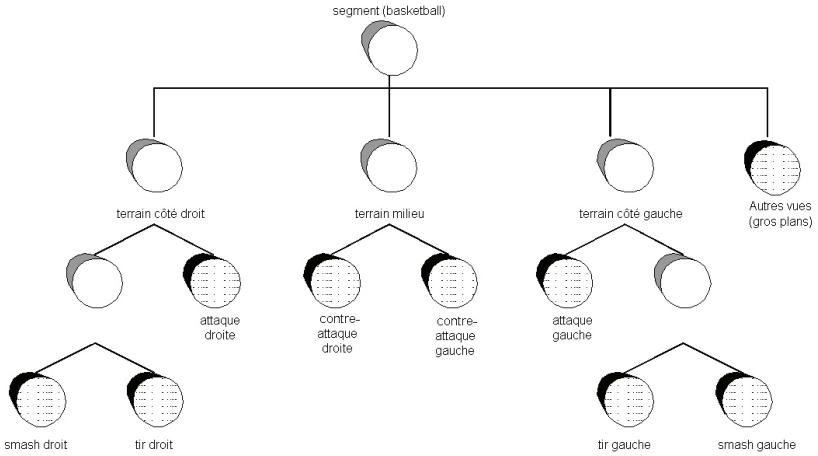


FIG. 1.4: Arbre de décision pour la classification d'un segment de basketball en 9 événements (feuilles en grisé).

Pour d'autres types d'événements, l'information temporelle exploitant les règles de production est de première importance. Par exemple, des règles de décision sur l'évolution du mouvement de la caméra dans une direction donnée permettent de détecter les attaques et les contre-attaques dans un jeu de basketball [47]. Les tirs au panier sont alors caractérisés par d'autres règles heuristiques sur la succession des plans comme : "un tir au panier est un segment de la vidéo contenant un zoom de la caméra ou un gros plan, juste après une contre-attaque" ou "un tir au panier a eu lieu lorsqu'il y a eu deux contre-attaques successives dans des directions opposées". Dans ce cas, deux événements ont d'abord été identifiés par une analyse de l'évolution des mouvements de la caméra, puis un troisième événement est déduit de l'entrelacement temporel des plans.

Cette notion d'entrelacement temporel est modélisable de façon plus formelle par des modèles logiques temporels. La figure 1.5 représente les modèles logiques utilisés pour détecter des tirs au but au football, à partir du mouvement de la caméra et de la localisation du jeu sur le terrain [1].

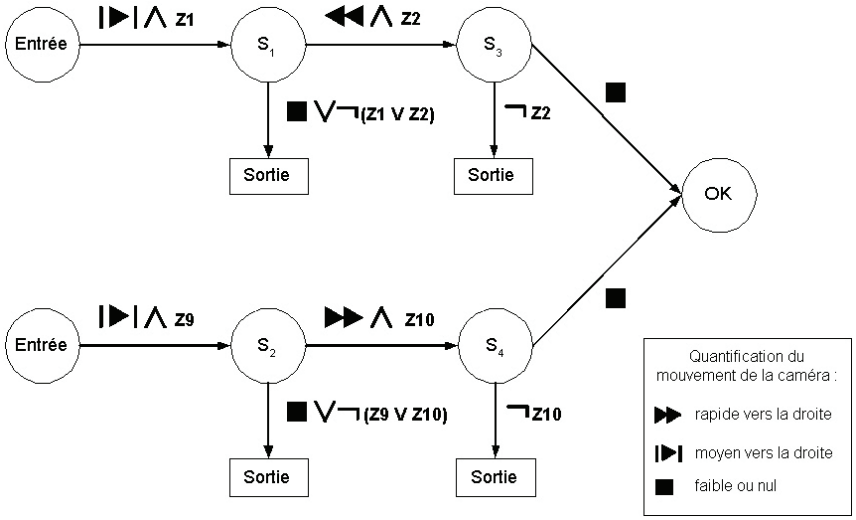


FIG. 1.5: Modèle logique temporel pour la détection des buts dans un match de football [1]. Sur les arcs : quantification du mouvement de la caméra et classification en zones du terrain (Z1...Z12) nécessaires à la transition entre états. La branche supérieure décrit un tir dans les cages de gauche, et la branche du bas dans les cages de droite. Si l'état *OK* est atteint, le tir est détecté.

D'autres approches consistent à d'abord détecter des événements de manière indirecte, en détectant (i) un ralenti [2, 32, 63], (ii) des mots-clés prédéfinis dans le flux textuel[1] associé à la vidéo [64], (iii) ou encore un changement du score dans les incrustations [57]. Dans le premier cas, les événements ne sont pas pré-identifiés comme but, tir au panier, etc. On sait seulement qu'une action a été jugée suffisamment importante pour être rediffusée. Cette hypothèse est particulièrement utilisée pour le football américain et le football, pour lesquelles la définition d'un événement important (hormis ceux donnant lieu à un score) est quelque peu subjective. C'est la subjectivité du producteur qui est exploitée ici. Dans les deux autres cas, les événements sont sémantiquement identifiés par le mot-clé, ou par le type de changement de score réalisé.

Une fois l'événement détecté, il est localisé dans la vidéo en analysant l'entrelacement temporel des plans. Ces approches en deux étapes réduisent dans un premier temps la fenêtre temporelle de recherche de l'événement. Lorsque l'événement est détecté par le biais des rediffusions, le modèle temporel *a priori* considéré pour la localisation de l'action, est représenté dans la figure 1.6. L'information *a priori* exploitée est qu'une rediffusion est insérée juste après que l'action a eu lieu, et n'est séparée d'elle qu'au plus d'un gros plan marquant la fin de l'action. Une fois la rediffusion détectée, l'action à laquelle elle correspond est donc recherchée dans les plans actifs la précédant immédiatement.

[1] close-caption stream

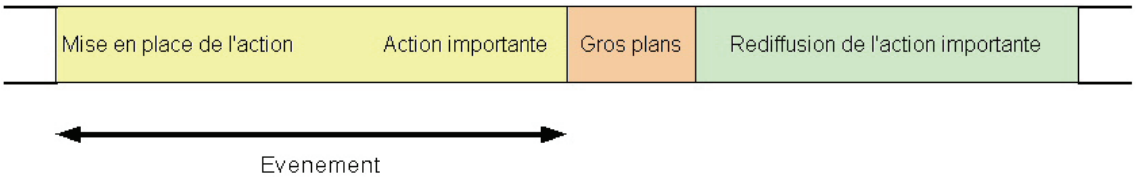


FIG. 1.6: Relation entre un événement important et sa rediffusion [2].

L'extraction de mots-clés pré-définis dans le flux textuel délimite une fenêtre temporelle de recherche de l'événement. Pour le localiser dans cette fenêtre, le modèle temporel de l'événement cherché dans [64] est simplement représenté par une séquence d'images exemple sélectionnée manuellement dans la séquence traitée. Les images des plans de la fenêtre temporelle de recherche sont alors comparées à la séquence modèle. Ce modèle temporel est bien entendu peu satisfaisant puisqu'il nécessite d'extraire une séquence exemple par événement recherché, ce qui rend le procédé semi-automatique.

La détection d'un événement par analyse du type de changement de score réalisé au baseball est donnée par un arbre de décision qui s'appuie sur les règles du baseball (Fig. 1.7) [57].

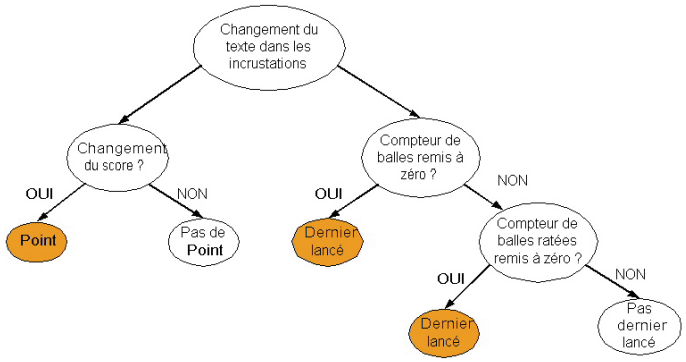


FIG. 1.7: Arbre de décision pour la détection d'un événement au baseball basée sur la reconnaissance du score incrusté.

La localisation de l'événement dans ce cas exploite une règle de production similaire à celle utilisée dans le cas de la détection des ralentis. Un segment contenant un événement est décomposé en une séquence d'éléments sémantiques : lancer, événement, vue non-active indiquant la fin de l'événement, rediffusion et changement du texte incrusté. L'événement est donc localisé en trouvant la vue du lancer la plus proche du changement de score.

1.3.2.2 Approches probabilistes

Une séquence d'images étant une série temporelle, les modèles de Markov cachés (HMMs) sont employés naturellement pour modéliser l'évolution des attributs d'une séquence d'images.

Dans le cadre de la détection d'événements, chaque événement est modélisé par un HMM. Pour chaque nouvelle séquence d'attributs, les modèles sont alors mis en compétition. Un événement est détecté si le modèle qui lui est associé fournit une vraisemblance supérieure à un certain seuil. Les états des HMMs sont suggérés par la connaissance *a priori*.

Les HMMs sont utilisés au niveau du plan ou d'une séquence de plans. Dans le premier cas, la séquence observée est la séquence des attributs extraits pour chaque image et le HMM modélise l'évolution des attributs au cours du plan. La modélisation de l'évolution des paramètres du mouvement de la caméra permet par exemple de décrire un penalty (Fig. 1.8), un coup-franc ou un corner au football [65].

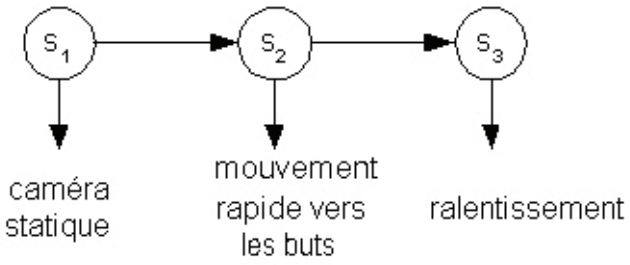


FIG. 1.8: Modèles de Markov cachés décrivant l'évolution des paramètres de mouvement dans un plan pour modéliser un penalty au football.

Au niveau d'une séquence de plans, ce sont les transitions entre les plans qui sont modélisées par des HMMs pour représenter des événements. La figure 1.9 représente les HMMs construits pour quatre événements du baseball. Les modèles reposent uniquement sur l'analyse des transitions entre les différents types de plans. La classification des plans est réalisée dans une étape préliminaire [44].

1.3.3 Segmentation en unités logiques

La segmentation en unités logiques consiste à trouver une description de la structure du document vidéo et à localiser les segments qui correspondent à cette structure. La structure est dite *dense*, lorsqu'elle décrit tous les segments de la vidéo, ou *partielle*, lorsqu'elle n'identifie qu'une partie des unités logiques.

Comme pour les événements, les unités logiques dépendent du type de sport considéré.

1.3.3.1 Structuration partielle

Certains sports sont caractérisés par une *vue fondamentale* qui indique le début d'une unité logique et donc les limites d'une structure de plus haut-niveau. Pour le tennis, la vue fondamentale est la vue de terrain correspondant au moment du service, pour le baseball, ce sera une scène de lancer, pour le football américain, la vue de l'engagement caractérisée par l'alignement des joueurs.

Zhong *et al.* [25, 41] formulent l'analyse de la structure comme le problème de détecter les vues fondamentales à l'aide d'un apprentissage supervisé et des règles spécifiques. La structuration se résume dans ce cas à rechercher les images de service du tennis, et les

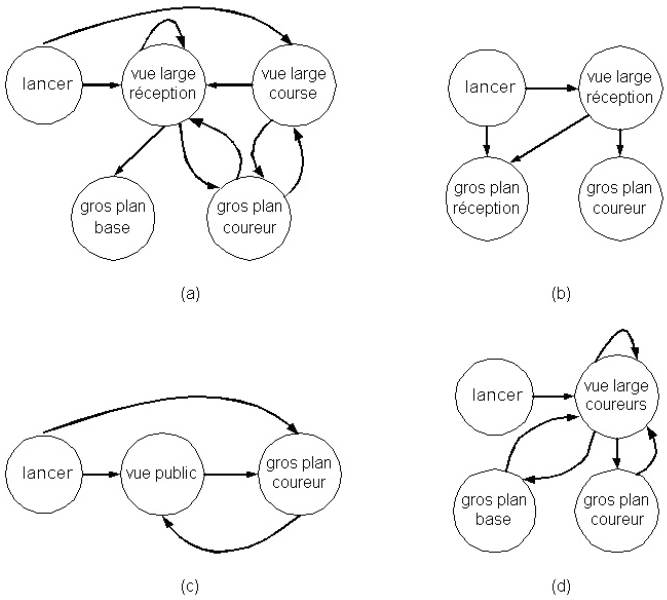


FIG. 1.9: Modèles de Markov cachés décrivant les transitions entre les plans pour modéliser dans une vidéo de baseball (a) une belle frappe, (b) une belle réception, (c) un home run, (d) un jeu dans le champ intérieur.

images de lancer du baseball, sans analyse d'une structure de plus haut niveau. Cette approche basée image n'intègre aucune information temporelle, pas même en terme de segmentation en plans.

De la même manière, mais en se basant sur une segmentation en plans de la vidéo, Lu *et al.* [45] ramènent le problème de la structuration à l'extraction des deux scènes majoritaires d'une séquence et à la classification des plans selon leur appartenance ou non à l'une ou l'autre de ces scènes. Les scènes de service au tennis et au volley sont déterminées par des règles heuristiques. Par exemple, pour le volley, le service est localisé au début d'un plan large contenant un mouvement de translation rapide.

Le baseball possède une structure bien déterminée représentée à la figure 1.10. L'unité logique à la base de la structure est le cycle lancer-frappe qui se compose de différents plans. Kawashima *et al.* [24] détectent les images représentant les lancers et les frappes de balle en les comparant avec des séquences modèles. Le modèle de la zone de score étant fourni pour la vidéo considérée, le changement de batteur est identifié par analyse du score incrusté. L'objectif est intéressant, même si seule une structuration partielle est réalisée, et si la mise en œuvre est rudimentaire.

1.3.3.2 Structuration dense

De façon générale, tous les sports peuvent être segmentés en phases de jeu et de non-jeu. Les phases de jeu sont définies par l'intervalle de temps où la balle est en jeu, et les phases de non-jeu correspondent aux moments où l'action est arrêtée : score, ballon en dehors de la surface de jeu, arrêts et préparation, ... La segmentation en phases de jeu et

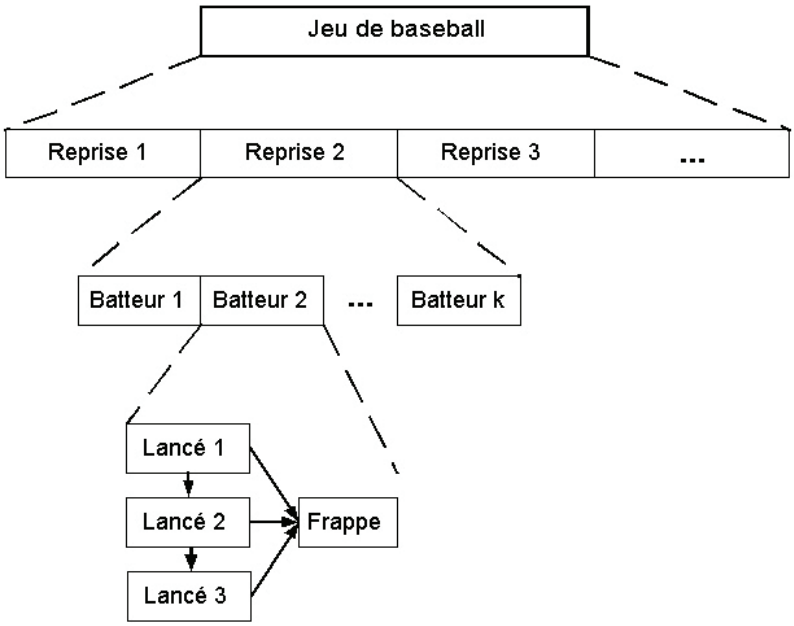


FIG. 1.10: Structure d'un match de baseball.

de non-jeu est une structuration dense mais pas de très haut niveau.

Approches déterministes

Une mise en application immédiate de la segmentation en phases de jeu et de non-jeu pour un match de football américain, consiste à rechercher une image caractéristique d'un début de jeu (vue du terrain dans laquelle les joueurs sont alignés), puis à rechercher l'image d'une scène de non-jeu suivante (gros plan notamment, dans lequel on ne voit pas le terrain) et la rupture de caméra associée indiquant la fin de l'action [28]. Cette méthode simple basée image est l'équivalent d'une méthode basée plan qui consisterait à détecter les plans contenant une vue fondamentale et à les labeliser comme jeu, tous les autres plans correspondants à du non-jeu.

Effectivement, une analyse simpliste est d'attribuer les phases de jeu aux vues globales représentant le jeu, et les phases de non-jeu aux gros plans. Cependant, la segmentation en phases de jeu et de non-jeu s'avère moins évidente qu'une classification des plans selon leur point de vue, lorsque les phases de jeu sont composées de plusieurs plans comme au football. Dans ce cas, il y a des vues globales ne contenant pas de jeu et des gros plans durant une phase de jeu. Il faut alors analyser le voisinage de chaque plan. Plusieurs approches réalisent cette analyse en se basant sur un ensemble de règles [31, 3, 63]. L'observation fondamentale est que les vues globales contenant effectivement une phase de jeu sont plus longues que les vues globales n'en contenant pas. Les séquences composées d'une succession de vues globales de longue durée sont immédiatement étiquetées comme "jeu". Puis, le voisinage de ces segments est étudié (Fig. 1.11). La segmentation est réalisée uniquement sur le type, la longueur et le voisinage de chaque plan. Cette méthode simple présente deux inconvénients : d'une part, elle nécessite de fixer des seuils qui peuvent varier d'une séquence à l'autre,

d'autre part, elle repose sur l'hypothèse que les plans contenant du jeu sont plus longs que ceux n'en contenant pas. Cette hypothèse n'est pas vérifiée dans le cas d'une action rapide, comme un but à la suite d'un corner pour le football ou un service gagnant au tennis.

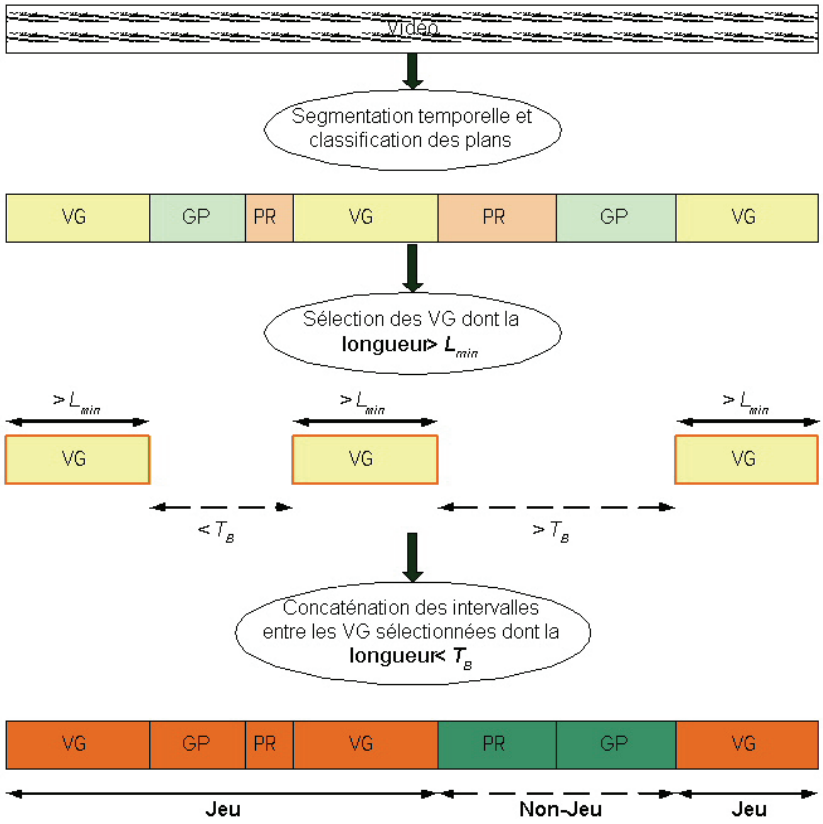


FIG. 1.11: Segmentation en phases de jeu et de non-jeu par des règles heuristiques [3]. VG : Vue Globale, GP : Gros Plan, PR : Plan Rapproché.

Approches probabilistes

Les systèmes d'analyse de la structure utilisent de plus en plus des modèles de Markov cachés, par exemple pour réaliser une segmentation en phase de jeu et de non-jeu d'une vidéo de baseball, ou de football américain [66, 67], ou de football [49]. Li *et al.* [67] proposent de modéliser les phases de jeu par un modèle de Markov à quatre états (Fig. 1.12). Ce modèle décrit les règles suivantes :

- une phase de jeu débute par une vue du lancer ;
- si le plan suivant représente une vue du terrain et présente un mouvement de caméra significatif, le jeu continue ;
- une phase de jeu se termine par un gros plan ou toute vue ne comportant pas de terrain ;
- l'intervalle de temps entre la fin de la phase de jeu, courante et le début de la prochaine phase de jeu est considéré comme non-jeu.

L'approche est basée image, bien que les frontières des phases de jeu coïncident avec celles des plans. Pour trouver les vues du lancer, chaque image est comparée à une image modèle d'une vue de lancer, ce qui suppose pour chaque nouvelle vidéo traitée d'en extraire une manuellement. Finalement, les phases de jeu sont les segments compris entre une vue du lancer et le gros plan qui la suit immédiatement. Tous les autres segments sont des phases de non-jeu. Pour un gros plan en cours d'action, par exemple sur le joueur qui rattrappe la balle, la phase de jeu est tronquée.

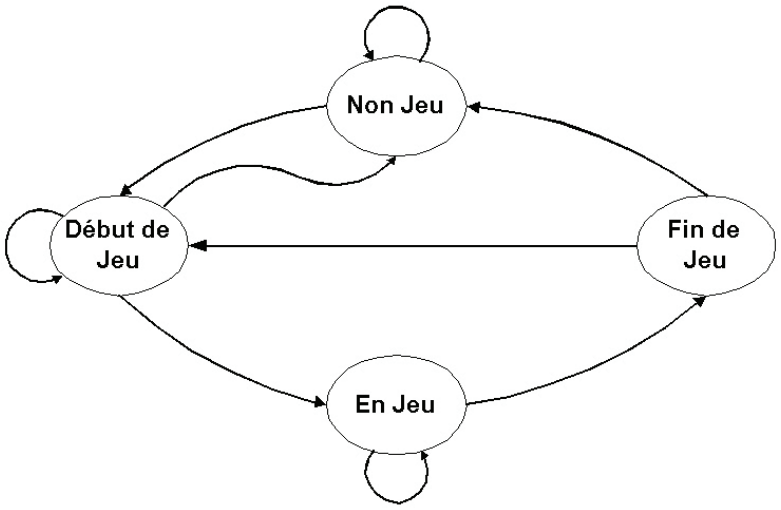


FIG. 1.12: Modèles de Markov à 4 états modélisant la phase "jeu" d'une vidéo de baseball ou de football américain.

Pour le football, l'approche est basée image car la segmentation en phases de jeu ou de non-jeu ne correspond pas forcément à la segmentation en plans. Chaque image est caractérisée par son taux de pixels de la couleur du terrain et l'intensité moyenne du mouvement. Six HMMs de topologies différentes sont construits respectivement pour les phases de jeu et de non-jeu [49]. L'approche est représentée à la figure 1.13. La séquence d'images est analysée par le biais d'une fenêtre temporelle glissante. Chaque segment de la fenêtre est mis en entrée des douze HMMs et les vraisemblances maximales pour chacune des classes jeu et non-jeu sont retenues. Un algorithme de programmation dynamique réalise alors la segmentation temporelle en fonction des probabilités de transitions entre les phases de jeu et de non-jeu et de la vraisemblance de chaque segment d'appartenir à la classe jeu ou non-jeu.

Nitta *et al.* [4] recherchent les unités logiques dans une vidéo de football américain. La structure d'un programme de sport est représentée comme une succession de scènes en direct, de rediffusions et de scènes en studio (Fig. 1.14). L'unité logique est alors définie comme l'intervalle de temps entre deux scènes de direct (*Live scene*).

La segmentation et la classification en scènes "Pré-Live", "Live", "Post-Live" et autres (cf. figure 1.14) est réalisée par réseau bayésien sur le flux textuel associé à la vidéo (*closed-caption stream*). Puis les segments textuels sont associés aux segments vidéo correspondants en cherchant les frontières les plus proches de la segmentation vidéo en plans.

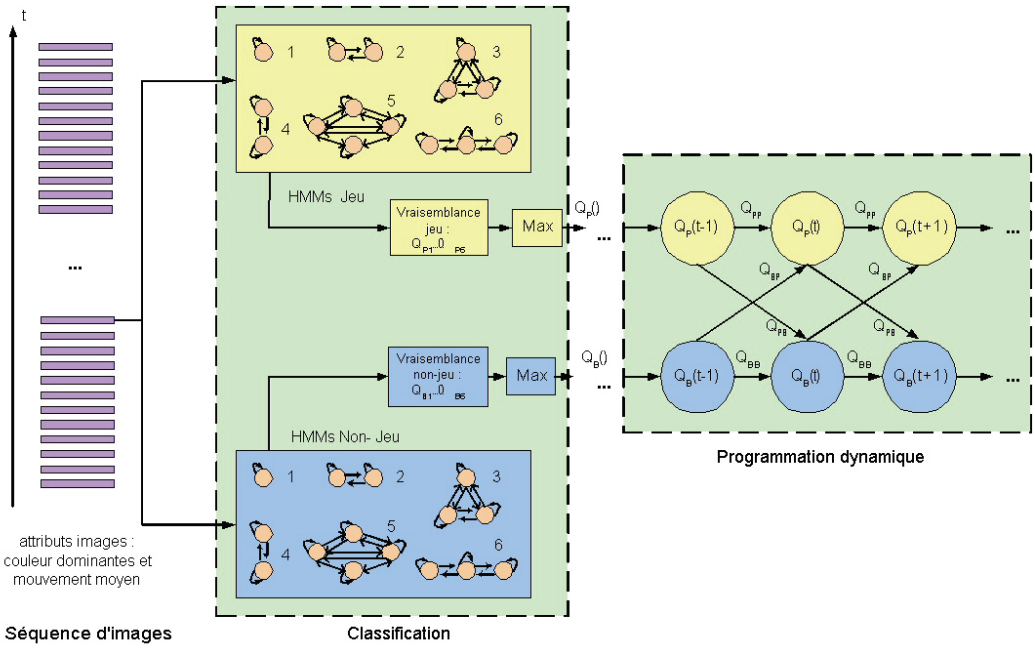


FIG. 1.13: Segmentation en phases de jeu et de non-jeu d'une vidéo de football par programmation dynamique.

Par ailleurs, la structure d'un match de football américain est bien déterminée. Elle peut être exprimée sous la forme d'un arbre. Le match est composé de la répétition d'un élément fondamental : "1rst Down" et "2nd Down" dans la figure 1.15. Cet élément est l'unité logique basique décrivant un programme de football américain.

La segmentation de la vidéo en unités logiques permettrait, dans une étape ultérieure, de les structurer au sein de l'arbre correspondant à la structure d'un match de football américain. Ce problème-ci n'est néanmoins pas traité.

1.4 Autre système spécifique : analyse des journaux télévisés

Nous avons présenté les systèmes spécifiques pour les retransmissions télévisées des événements sportifs. Il existe bien entendu de nombreux autres systèmes spécifiques. Le domaine le plus proche au niveau des objectifs et des méthodes mises en œuvre est celui de l'analyse des journaux télévisés. A l'instar de certains sports, les journaux télévisés ont une structure bien déterminée, indiquée à la figure 1.16. L'objectif est généralement de générer une table des matières [18, 68] et des résumés [69].

Un journal télévisé est composé d'un certain nombre de sujets, dont les titres sont d'abord annoncés. Chaque sujet, relatif à l'actualité, la politique, le sport, etc est ensuite introduit par le présentateur et développé sous forme de reportage. Le journal s'achève souvent avec l'interview sur le plateau d'un invité, puis se conclut avec le bulletin météorologique.

Les journaux télévisés sont tous produits de la même façon selon des règles de pro-

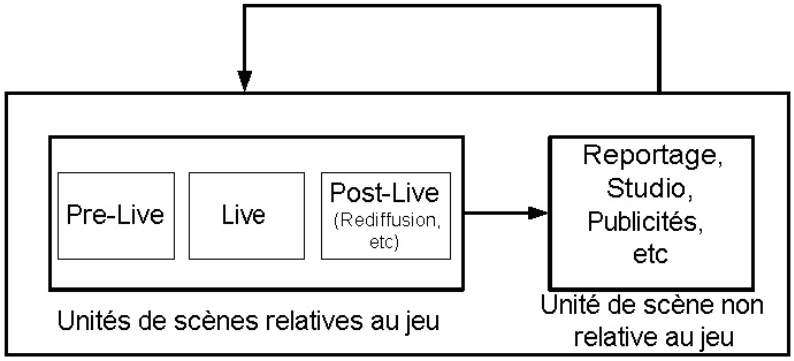


FIG. 1.14: Structure d'un programme de sport télévisé [4].

duction relativement rigides. De façon générale, un journal débute par un plan d'ensemble du plateau durant lequel le présentateur annonce les titres du journal qui seront développés. Dans le reste du document, les reportages et les scènes de plateau dans lesquelles le présentateur annonce le sujet suivant se succèdent comme la figure 1.17 l'illustre.

Les unités logiques d'un journal télévisé sont naturellement définies par les différents sujets abordés. Les scènes de plateau sont au journal télévisé ce que les vues du terrain sont aux émissions sportives. Elles constituent les *vues fondamentales* indiquant le début d'une unité logique et les limites d'une structure de plus haut niveau.

Une étape fondamentale de l'analyse d'un journal télévisé consiste donc à détecter les vues du présentateur. Ces vues partagent de nombreuses caractéristiques qui facilitent leur détection. D'abord, ces plans sont filmés par une caméra statique, ce qui implique que le présentateur sera quasiment toujours localisé au même endroit dans les images. Ainsi, la comparaison de la différence moyenne entre images successives de la région présumée du présentateur est utilisée pour détecter ces plans dans l'article [18].

La structure spatiale de ces images est identique tout au long de la séquence. De plus, l'unité de lieu du plateau assure un contenu visuel de ces vues similaire pour tous les plans du présentateur d'une même vidéo : même arrière-plan, même présentateur au premier plan. La détection des plans du présentateur repose souvent sur la détection et la reconnaissance de visages [70, 71]. La connaissance *a priori* de la structure spatiale d'un plan du présentateur (Fig. 1.18) guide la détection de visages, en restreignant la position et en limitant la taille des visages à détecter.

Une autre caractéristique utilisée pour rechercher ces vues est la fréquence d'apparition de ces plans et leur similarité [72]. Les plans sont regroupés par clustering et le plus large groupe de plans similaires est étiqueté comme vue du présentateur.

La recherche des zones d'incrustation et la détection du texte sont également utilisées. La détection d'un titre incrusté permet d'identifier un plan introduisant un nouveau sujet.

Nous avons vu pour le sport que les plans étaient distingués par leurs classes selon qu'ils sont des vues larges du terrain, des plans moyens ou des gros plans. Les vues larges sont celles qui ont le plus d'importance et la classification de ces plans est généralement réalisée à partir de la taille de la zone de terrain dans l'image. Pour les journaux télévisés, on retrouve le même type de classes, mais les plans d'intérêts sont au contraire les plans

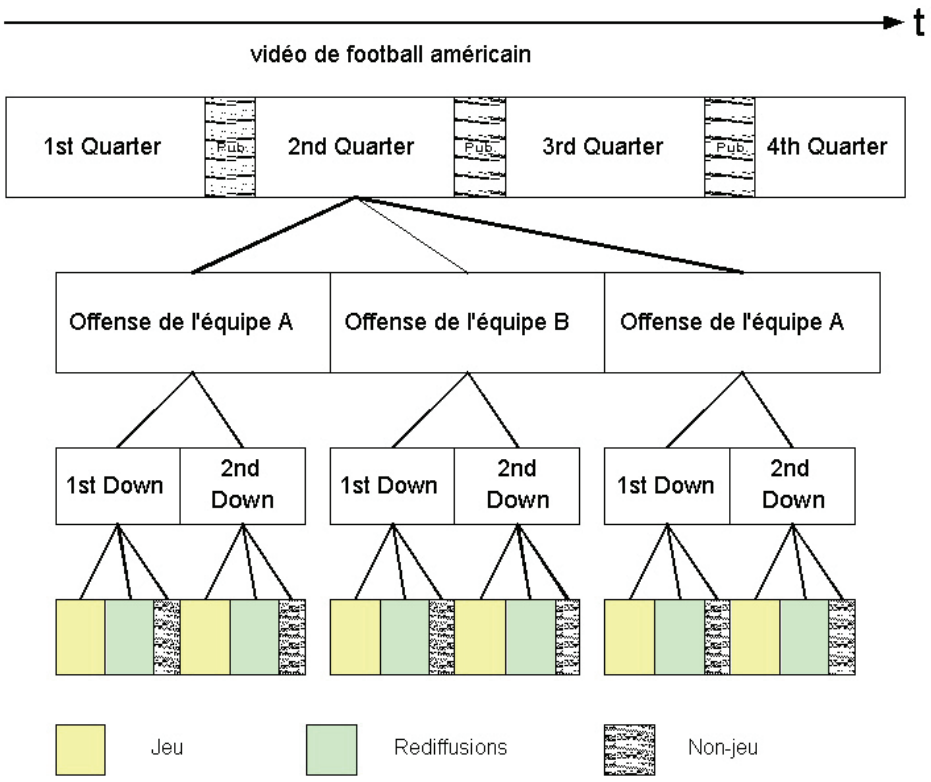


FIG. 1.15: Structure d'une vidéo d'un match de football américain [4].

moyens et les gros plans. La classification des plans est alors réalisée selon la taille du visage détecté dans l'image [73]. Chaisorn *et al.* [74] proposent une classification des plans plus sémantique en treize catégories comme présentateur, interview, reportage en direct, sport, météo, finance, publicités, etc.

La détection des vues fondamentales et la classification des plans donne un étiquetage simple des plans en vue "présentateur" ou "reportage" [72].

Une fois les vues du présentateur identifiées, la connaissance *a priori* de la structure du document va permettre de le segmenter en unités logiques. Des règles logiques de décision construisent, par exemple, des graphes de relations entre les prises de vues reflétant la structure relationnelle du document [75]. Tout comme pour les événements sportifs, les méthodes de segmentation en unités logiques sont basées sur des règles heuristiques exploitant les règles de production [76], ou sur des modélisation de la structure temporelle du journal télévisé, diagramme d'états [16] ou modèles de Markov cachés [74, 77]. Il s'agit bien souvent d'une segmentation en scènes de plateau et reportages. Pour une analyse plus précise, d'autres unités logiques peuvent être introduites, telles que les interviews et la météo [77](Fig. 1.19).

Dans d'autres approches, la segmentation en scènes s'appuie sur les frontières extraites des flux audio [68] ou textuels [78, 79].

Les scènes de dialogue sont généralement identifiées par le motif temporel bien particulier qu'elles exhibent : si A et B sont deux gros plans de deux personnes différentes, un

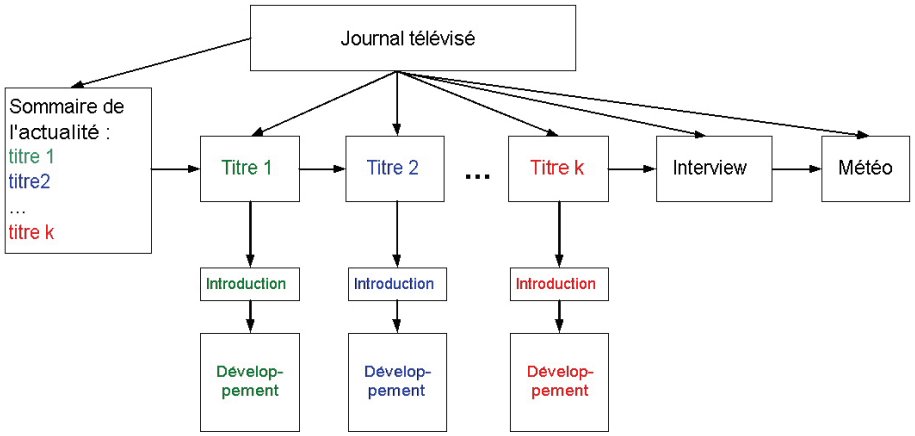


FIG. 1.16: Structure d'un programme de journal d'information.

dialogue est caractérisé par l'alternance des plans A et B, appelés champ et contre-champ dans le domaine cinématographique. Pour détecter les dialogues, des scènes vérifiant le motif $ABABABABAB$ sont cherchées. La détection des scènes de dialogues est utilisée pour l'analyse des journaux télévisés, mais également des films [80]. C'est en effet l'une des rares structure qu'il est possible d'identifier dans un contenu aussi hétérogène que celui d'un film. Là encore, la détection des scènes repose sur un ensemble de règles ou sur une modélisation du motif temporel par HMM [73, 81].

D'autres exemples d'émissions structurées sont les jeux télévisés et les "talk shows", dans lesquels l'animateur et les invités doivent être identifiés [82].

L'analyse des journaux télévisés et des événements sportifs partagent donc la même exploitation de la connaissance *a priori* de la structure spatiale et temporelle des vidéos. Les méthodes mises en œuvre, basées modèle, sont relativement identiques.

1.5 Discussion

Nous venons de passer en revue un ensemble non exhaustif des méthodes existantes pour l'analyse spécifique des vidéos de sport et des journaux télévisés. Ces systèmes spécifiques utilisent intensivement la connaissance *a priori*. Comme toutes les méthodes fondées sur l'utilisation d'informations *a priori*, cela requiert de mener au préalable une analyse formelle des documents étudiés. Nous avons montré quels types d'informations étaient utilisés et différentes méthodes pour les exploiter depuis l'analyse bas-niveau jusqu'à l'interprétation sémantique. De façon générale, la prise en compte des règles de production entraîne une dépendance des méthodes vis-à-vis des artefacts des producteurs (*i.e.* manquements à la règle). La figure 1.20 récapitule de manière schématique les différents niveaux d'analyse et la connaissance *a priori* exploitée à chaque niveau.

Les attributs de mouvement sont plus utilisés pour la détection d'événements dans les sports comme le football, le basketball, le cricket ou le baseball, dont les plans présentent de grands mouvements de caméra. Les plus grands mouvements de la caméra impliquent

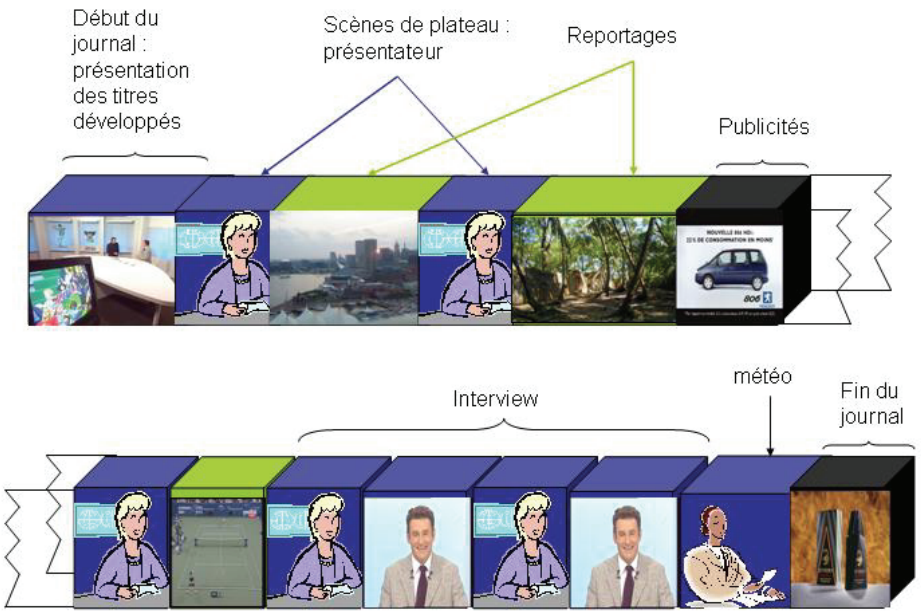


FIG. 1.17: Règles de production et structure temporelle d'un programme de journal d'informations. Scènes de plateau en bleu, reportages en vert.

aussi que le contenu d'un plan change énormément. Pour ces raisons, les approches basées images sont préférées aux approches basées plans pour l'analyse de ces sports.

Les caractéristiques objet comme la détection de visages, des joueurs ou du ballon sont plus précises pour analyser une action que des attributs globaux, mais trop coûteuses en temps de calcul et pas assez robustes. En effet le nombre important de joueurs et leur trajectoire aléatoire rendent leur suivi très difficile, de même que la très petite taille et la vitesse de déplacement de la balle au tennis rendent son suivi quasiment impossible dans des conditions acceptables de temps de calcul.

1.5.1 Les différentes approches : déterministes ou probabilistes

Dans les approches déterministes, le raisonnement se base sur un ensemble de règles spécifiques au domaine et se traduit par un ensemble de règles de décisions. Ces approches sont faciles à implémenter, rapides à calculer et donnent de bons résultats [24, 25]. Cependant, les règles de décision ne sont pas toujours simples à formuler explicitement, en particulier lorsqu'elles reposent sur un nombre important d'indices. La fixation de seuils est souvent une tâche difficile et donc, un autre inconvénient de ces approches. Sans compter qu'un seuil fixe ne couvre pas toutes les variations que l'on trouve dans les vidéos.

Les modèles d'inférence probabilistes, tels que les modèles de Markov cachés, permettent d'intégrer des règles plus complexes. De plus, les incertitudes liées à l'analyse de la vidéo sont intégrées au modèle : il n'y a pas de seuils difficiles à choisir. En contrepartie,

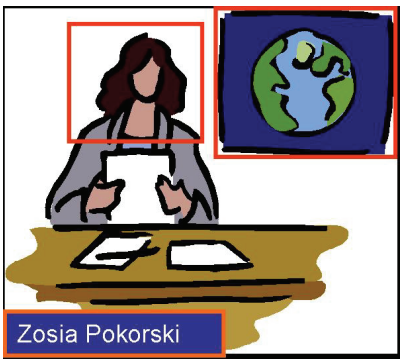


FIG. 1.18: Structure spatiale d'un plan du présentateur. En rouge, les objets d'intérêt pouvant être extraits pour l'analyse de la séquence.

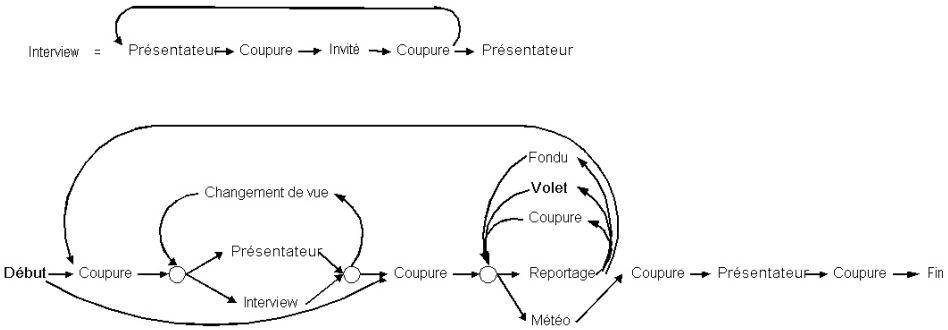


FIG. 1.19: Modélisation de la structure temporelle d'un journal télévisé par modèles de Markov cachés.

ces modèles nécessitent la mise en œuvre d'une phase d'apprentissage.

Li *et al.* ont comparé les approches déterministes et probabilistes pour la détection des phases de jeu dans une vidéo de baseball [67] et une vidéo d'entraînement de football américain [66]. Dans les deux cas, les résultats obtenus par des règles de décisions sont très bons. Dans le cadre de l'environnement très contraint des vidéos d'entraînements, dont la structure est aussi bien déterminée que celle d'un journal télévisé, l'approche probabiliste donne encore de meilleurs résultats. En revanche, pour la vidéo de baseball, les modèles de Markov cachés sont supplantés par les règles de décision. Mais leur modèle était trop simple et par conséquent peu adapté.

1.5.2 Indexation et structuration

Nous avons distingué deux formes d'analyse ayant des objectifs différents : la détection d'événement et la structuration. La détection d'événement identifie des occurrences de certains concepts sémantiques apparaissant de façon plus ou moins ponctuelle dans la vidéo. La structure représente le niveau de composition syntaxique du contenu de la vidéo. Le processus de structuration détermine et localise des unités logiques composant le document.

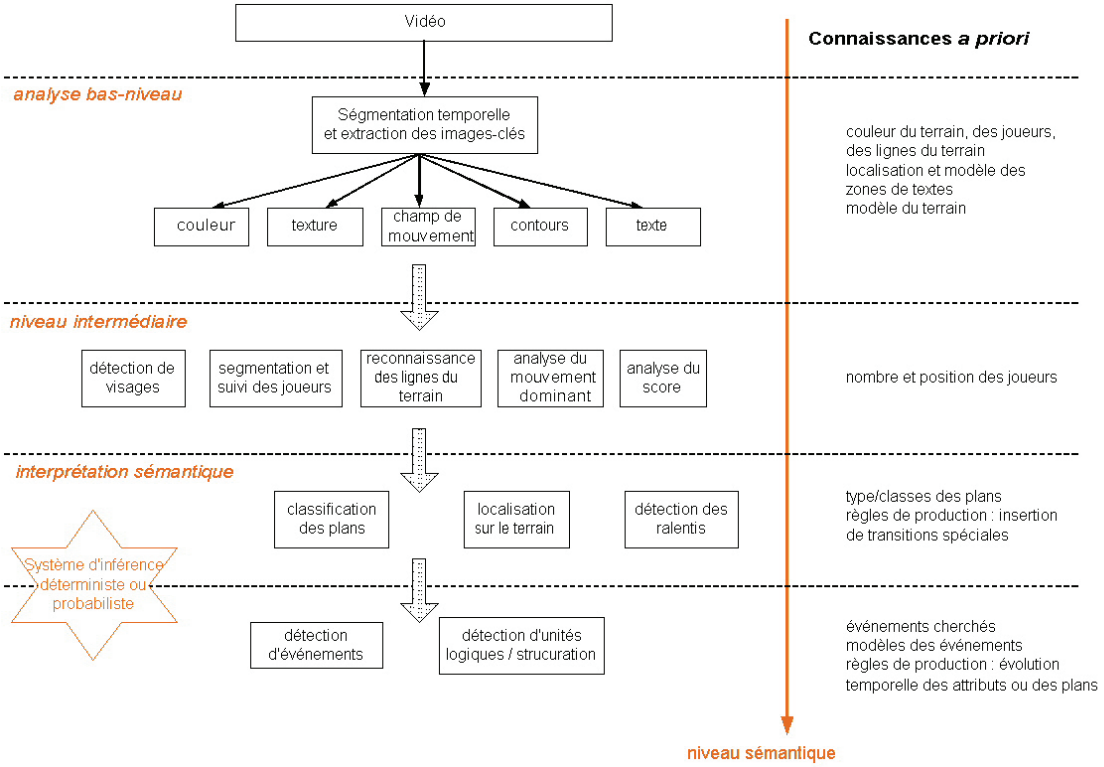


FIG. 1.20: L'apport des connaissances *a priori* sur les différents niveaux de l'analyse des vidéos de sport.

La détection d'événements construit une table des index de la vidéo, quand la structuration élabore une table des matières. Ces deux approches sont donc complémentaires.

Li *et al.* [2] modélisent tous les sports en utilisant le concept d'événement selon la remarque suivante : la plupart des sports sont modélisables comme une simple concaténation de segments d'"événements" et de "non-événements" (Fig. 1.21), un "événement" étant défini comme un segment (pouvant contenir plusieurs plans) durant lequel une action importante se joue. L'objectif de l'analyse est alors de définir tous les événements d'une vidéo à l'aide d'informations *a priori*, puis de les détecter.

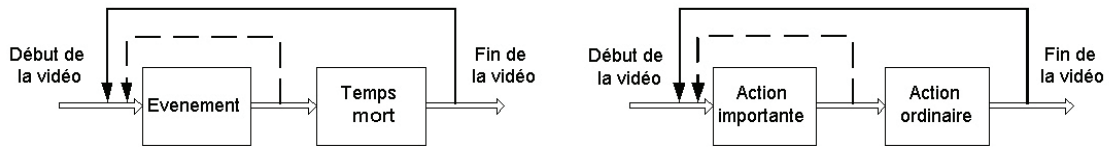


FIG. 1.21: Modèle général d'une vidéo de sport : à gauche pour les sports à action discontinue, à droite pour les sports à action continue.

Nous voyons que la notion d'événement est fortement corrélée à la notion d'événement

"intéressant". La détection d'événement se heurte à la difficulté d'évaluer ce qu'est un événement intéressant, hormis quelques évidences comme les buts au football. C'est la raison pour laquelle certaines approches se basent sur les événements généralement montrés dans les magazines de sport, ou prises en compte dans les statistiques sur le match. Une autre façon est de considérer qu'un événement est important s'il est rediffusé. La détection des événements est alors ramenée à la détection des rediffusions [2].

La structuration repose sur l'analyse de l'entrelacement temporel des plans. Peu de travaux traitent de la structuration dense d'une vidéo, et lorsqu'elle est abordée, la structure analysée n'est pas de très haut niveau. Il s'agit d'un découpage de la vidéo en scènes de jeu et de non-jeu, les scènes de jeu étant définies comme un ensemble de plans consécutifs pendant lesquels le jeu évolue, et les scènes de non-jeu comme l'ensemble des plans consécutifs pendant lesquels la balle est hors-jeu ou les joueurs se préparent.

Nous pensons que, pour un sport donné, une sémantique plus riche peut être formée sur la base de ces deux phases.

1.5.3 Les différents types de sports

Les sports diffèrent tellement dans leur nature qu'il est inévitable que l'analyse de vidéos soit spécifique à un sport donné, ou au mieux, à quelques sports sémantiquement similaires.

Nous néanmoins distinguons deux grands types de sports : les sports *contraints par le temps* (ou encore sports *à temps borné*) et les sports *contraints par le score* (ou encore sports *à score borné*). Les premiers sont des sports dont la durée de jeu est prédéterminée comme le football ou le basketball. Durant le temps imparti, éventuellement divisé en manches, les équipes marquent le plus de points possible. Il est impossible de prévoir combien de points seront marqués durant un match, ni quand ils le seront. Ces événements apparaissent de façon ponctuelle et imprévisible. Pour la seconde catégorie, le temps de jeu n'est pas déterminé. C'est le nombre de points que doivent marquer les parties opposantes pour remporter la victoire qui est fixé. C'est le cas par exemple du tennis ou du baseball.

Dans tous les sports, l'action est continue jusqu'à ce que la balle soit hors-jeu ou qu'un point soit marqué. Cependant pour les sports à score borné, les points sont des événements qui arrivent régulièrement. Le déroulement du jeu est caractérisé par des arrêts fréquents de l'action, ce qui rend finalement cette dernière plutôt discontinue.

La détection d'un événement du type point marqué, de par son occurrence régulière dans les sports à score borné ne présente pas le même caractère remarquable que pour un sport à temps borné. En revanche, les sports contraints par le score présentent souvent une structure hiérarchique déterminée par le comptage des points :

- un match de baseball peut se découper en 9 manches, chaque manche étant caractérisée par le passage de 3 à 9 joueurs à la batte, et chaque joueur à la batte ayant droit à 3 lancers (cf. Fig. B.3) ;
- un match de tennis se décompose en 3 à 5 sets, chaque set étant composé d'au moins 6 jeux, et chaque jeu se décomposant lui-même en points, chaque joueur ayant droit à 2 services pour engager le point.

A l'inverse, la structure des sports à temps borné est beaucoup plus pauvre : un match de football se décompose en 2 mi-temps de 45mn. Les joueurs essayent de marquer le plus de buts possibles durant chacune de ces mi-temps.

Pour toutes ces raisons, nous estimons que la détection d'événements est plus adaptée à l'analyse des sports *contraints par le temps* dont la description de la structure ne fournit pas d'informations intéressantes pour une navigation non-linéaire dans le document. Symétriquement, *les sports à score borné* qui présentent une forte structure riche en informations, se prêtent d'avantage à l'analyse de la structure.

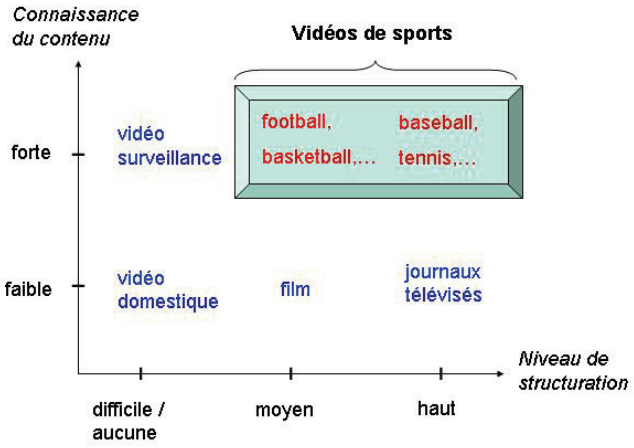


FIG. 1.22: Contexte de l'analyse des vidéos de sports.

1.6 Conclusion

Dans ce chapitre, nous avons proposé un panorama des méthodes existantes pour l'analyse spécifique des vidéos de sport, et nous avons fait le parallèle avec l'analyse des journaux télévisés. Nous avons distingué deux formes de connaissance *a priori* exploitée par ces systèmes : la connaissance liée au contenu de la vidéo (contextuelle), et la connaissance liée aux règles de production utilisées pour la télédiffusion. Nous avons montré comment ces informations étaient utilisées depuis l'analyse bas-niveau jusqu'à l'interprétation sémantique des documents. La plupart des systèmes s'intéressent à la détection d'événements importants, qui, concaténés, permettent de créer un index ou un résumé de la vidéo. Peu de travaux traitent de la structuration des vidéos, et encore moins d'une structuration dense. La structure analysée n'est alors pas de très haut niveau : il s'agit juste d'un découpage de la vidéo en scènes de jeu et de non-jeu.

La spécificité de chaque sport rend une approche générique des vidéos de sport difficile. Pour atteindre un haut-niveau de compréhension sémantique, la connaissance *a priori* est exploitée à chaque étape du système. De plus, les concepts même d'événements intéressants et de structure sont variables selon la nature du sport traité. Nous avons distingué les sports *contraints par le temps* et les sports *contraints par le score*, et nous avons expliqué pourquoi, de notre point de vue, nous estimons que la détection d'événements est plus adaptée à

l'analyse des sports *contraints par le temps* et la structuration aux *sports à scores bornés*. Nous pensons que, pour cette dernière famille de sport, une sémantique plus riche que la notion de "jeu" et de "non-jeu" peut être formée sur la base de ces "événements".

En conclusion, la structuration est un domaine de l'analyse des vidéos de sports encore récent. Les approches semblent aujourd'hui converger vers l'utilisation des modèles de Markov cachés. Le but du chapitre qui suit est d'apporter une contribution pour la structuration des vidéos possédant une structure complexe mais bien définie.

Chapitre 2

Représentation des unités logiques par modèles de Markov cachés

2.1 Introduction

2.1.1 Structuration d'un document

On appelle structure d'un document vidéo la structure syntaxique, sémantique et hiérarchique selon laquelle les images sont organisées au sein d'un document vidéo.

- La syntaxe se rapporte à l'aspect formel d'un langage, et à ses règles d'écriture (par analogie avec la syntaxe d'une phrase). Dans le cas d'un document vidéo, la structure syntaxique recouvre toutes les règles formelles de construction du document, dont un exemple est l'alternance prise de vue-transition ;
- Par sémantique, on entend tout ce qui est relatif au sens et à la signification des unités composant le document ;
- La structure hiérarchique rend compte de l'organisation des unités extraites du document en plusieurs niveaux hiérarchiques les uns par rapport aux autres, tant du point de vue syntaxique que sémantique.

La structure d'une séquence vidéo n'est pas directement accessible : elle nécessite d'être extraite par une série de traitements constituant le processus de structuration d'un document vidéo. Avant d'entamer le processus de structuration, il est nécessaire d'identifier les unités lexicales (éléments structurants) qu'on appellera *unités logiques* et qui sont les unités de base dont l'arrangement constitue la structure.

2.1.2 Présentation des sports analysés

Nous avons choisi deux sports comme cas d'étude : le tennis et le baseball. Ils ont été choisis en tant que de sports à scénario, dont le déroulement est régi par des contraintes de scores (voir 1.5). Ils présentent une structure bien déterminée, complexe et hiérarchique. Nous ne fournirons cependant de résultats expérimentaux que pour le tennis, faute de données. Le cas du baseball illustre l'application de notre approche à un autre sport, soulignant par là les aspects communs exploités de cette famille de sports.

Les unités logiques définies dans ce chapitre résultent de l'analyse syntaxique, sémantique et hiérarchique des vidéos de sport. Cette analyse se base sur deux formes d'infor-

mations *a priori* :

1. les règles intrinsèques du sport considéré dont dérivent les structures sémantique et hiérarchique, et nombre d'informations *a priori*. Les informations exploitées sont diverses : nombre et position des joueurs, modèle du terrain, déroulement et structure du jeu.
2. les règles de production dont dérive la structure syntaxique. La syntaxe est définie comme l'ensemble des structures et des règles qui permettent de produire tous les énoncés appartenant à une langue et seulement eux. Dans le cadre de la retransmission télévisée, les événements sportifs sont généralement soumis à des règles de réalisation spécifiques. Ces règles résultent :
   - du nombre fini de caméras utilisées pour retransmettre un événement sportif ;
   - de leur position souvent fixe et caractéristique de l'événement filmé ;
   - de leur utilisation : à un instant donné, le point de vue fournissant l'information la plus pertinente est sélectionné par le réalisateur.

Les diffusions télévisuelles sportives sont composées de scènes caractéristiques produisant des motifs répétitifs en raison des règles de réalisation. Dans la suite de ce document, on entend par unités logiques l'ensemble de ces scènes caractéristiques, et par syntaxe les règles régissant leur arrangement afin de produire des motifs répétitifs.

2.1.3 Méthode proposée

Dans ce chapitre, nous présentons une méthode de segmentation en unités logiques d'une vidéo de sport en utilisant des indices visuels. Cette méthode repose sur une modélisation statistique de l'entrelacement temporel des plans de la vidéo. Le cadre général de la modélisation est celui des modèles de Markov cachés. Les indices visuels sont utilisés pour caractériser le type des plans. Chaque unité logique est représentée par un modèle de Markov caché, intégrant les informations *a priori* sur le contenu de la vidéo, ainsi que sur les règles de production.

Tout d'abord, une segmentation vidéo est réalisée en détectant les transitions abruptes et graduelles entre les plans vidéos. Une image-clé représentative du contenu du plan et ses attributs vidéos sont extraits de chaque plan vidéo obtenu. La séquence des attributs vidéo des plans forme la suite d'observations qui est décodée par un processus de modèles de Markov cachés (Fig. 2.1). Le but du décodage est d'associer un état à chaque observation. La connaissance de cet état et du modèle auquel il appartient permet d'identifier les unités logiques.

Ce chapitre est construit de la manière suivante :

- dans une première partie 2.2, nous rappelons le formalisme des modèles de Markov cachés ;
- dans une deuxième partie 2.3, nous réalisons une analyse formelle des documents vidéo de tennis et nous présentons les unités logiques que nous avons définies pour ce sport. Nous détaillons ensuite leur modélisation par modèles de Markov cachés ;
- en 2.4, la segmentation temporelle de la vidéo et la caractérisation des plans, qui mènent à l'obtention des observations, sont présentées ;
- la même approche est illustrée dans le cadre du baseball dans la partie 2.5 ;
- enfin la dernière partie 2.6 est consacrée aux résultats expérimentaux.

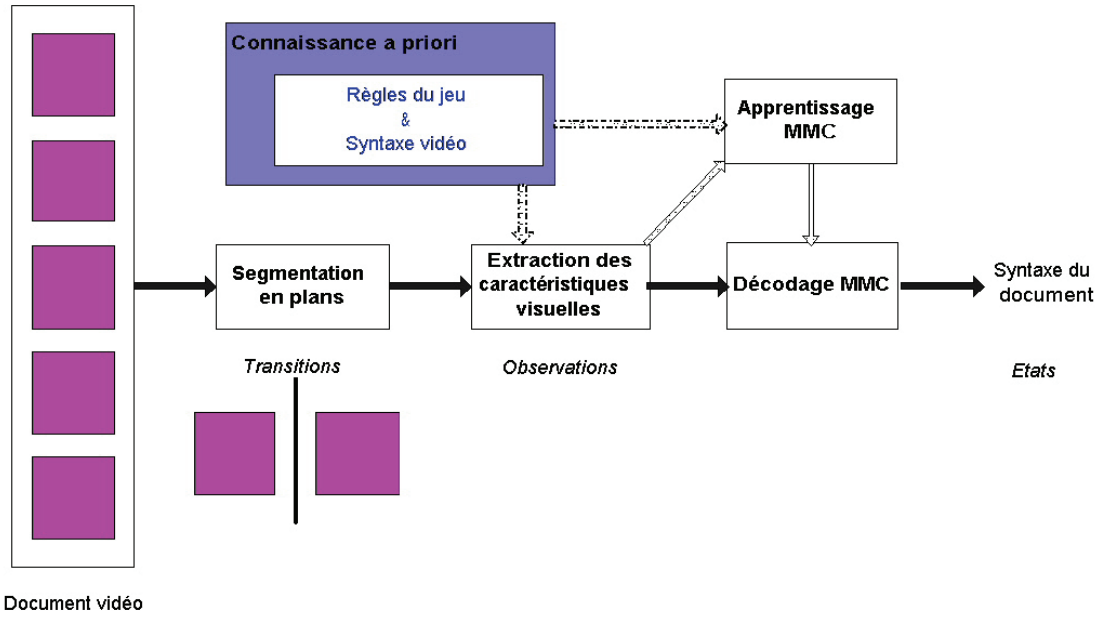


FIG. 2.1: Processus de structuration de la vidéo.

2.2 Modèles de Markov cachés

Ce paragraphe rappelle les notions fondamentales des Modèles de Markov Cachés (ou HMMs pour *Hidden Markov Models*) nécessaires à la compréhension de la suite de ce document. Pour une introduction plus approfondie, on se reportera au tutoriel de Rabiner [83].

2.2.1 Chaîne de Markov à états discrets

L'hypothèse sous-jacente des HMMs est que le signal étudié est défini comme un processus stochastique dont les paramètres sont estimés d'une façon précise et bien définie. Les HMMs modélisent l'évolution dans le temps de ce processus stochastique.

Le système étudié est décrit comme étant, à chaque instant, dans un état parmi un nombre fini N d'états distincts notés $s_1, s_2, ..., s_N$. A intervalle de temps régulier, le système évolue et change ou non d'état.

L'état du système à l'instant t est noté q_t. Le passage d'un état s_i à un état s_j se fait avec une certaine probabilité :

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k, ...) \tag{2.1}$$

qui dépend des états antérieurs du système.

On parle de modèle de Markov du *premier ordre* lorsque l'état du système considéré à l'instant t ne dépend plus que de l'état de ce système à l'instant $t-1$:

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k, ...) = P(q_t = s_j | q_{t-1} = s_i) \tag{2.2}$$

Sous cette condition et en considérant que le modèle de Markov est homogène, on construit une matrice A des probabilités de transitions indépendante du temps :

$$A = (a_{ij})_{1 \leq i,j \leq N} \quad \text{telle que} \quad a_{ij} = P(q_t = s_j | q_{t-1} = s_i) \tag{2.3}$$

La matrice A est stochastique, *i.e.* :

$$\begin{cases} a_{ij} \geq 0 & \forall i, j = 1...N \\ \sum_{j=1}^{N} a_{ij} = 1 & \forall i = 1...N \end{cases} \tag{2.4}$$

Afin de spécifier les conditions initiales du système, on introduit un vecteur Π qui fournit la distribution de probabilité de cet état :

$$\Pi = (\pi_i)_{1 \leq i \leq N} \quad \text{telle que} \quad \pi_i = P(q_1 = s_i) \tag{2.5}$$

et qui possède également la propriété :

$$\begin{cases} \pi_i \geq 0 & \forall i = 1...N \\ \sum_{i=1}^{N} \pi_i = 1 \end{cases} \tag{2.6}$$

2.2.2 Modèles de Markov Cachés

Dans le paragraphe précédent, l'état du système est déterminé par l'observateur, autrement dit chaque état correspond à un seul phénomène physique. On parle de modèles de Markov *observables*. Ces modèles ne sont pas assez puissants pour traiter le cas de distributions complexes. En effet, que se passe-t-il lorsque l'observation faite ne permet plus de spécifier l'état du système ?

On parle alors d'états *cachés*, c'est-à-dire qu'on ne dispose plus que d'une probabilité de correspondance d'une observation à un état.

Soit M le nombre d'observations qu'il est possible de faire, et $V = \{v_1, v_2, ..., v_M\}$, l'ensemble discret des M symboles d'observations. On définit une matrice B des probabilités d'observation par :

$$B = (b_j(k))_{1 \leq j \leq N, 1 \leq k \leq M} \quad \text{telle que} \quad b_j(k) = P(o_t = v_k | q_t = s_j) \tag{2.7}$$

où o_t est l'observation faite à l'instant t, $v_k \in V$ un symbole observé et $b_j(k)$ est la probabilité de faire l'observation du symbole v_k en étant dans l'état s_j.

La matrice B est aussi stochastique, *i.e.* :

$$\begin{cases} b_j(k) \geq 0 & \forall j = 1...N, \forall k = 1...M \\ \sum_{k=1}^{M} b_j(k) = 1 & \forall j = 1...N \end{cases} \tag{2.8}$$

Un système modélisé par un HMM génère une séquence $O = o_1, o_2, ..., o_T$ de symboles observables, appelée *séquence d'observations* de longueur T. Les observations sont de plus supposées indépendantes sachant les états cachés $Q = q_1, q_2, ..., q_T$, ce qui s'écrit :

$$P(O|Q) = P(o_1, \ldots, o_T|q_1, \ldots, q_T) = \prod_{t=1}^{T} P(o_t|q_t) \tag{2.9}$$

Un modèle de Markov caché λ est défini par le triplet :

$$\lambda = (A, B, \Pi) \tag{2.10}$$

2.2.3 Problèmes fondamentaux

Pour qu'un modèle de Markov caché soit utilisé efficacement en pratique, il faut résoudre les trois problèmes de base suivants.

1. **Evaluation ou estimation de la probabilité d'observation.** Étant donnée une séquence d'observations O et un modèle λ, quelle est la probabilité $P(O|\lambda)$ de générer O par le modèle λ ?

   La probabilité de O sachant un modèle λ est la somme sur tous les chemins d'états Q des probabilités conjointes de O et de Q par rapport à ce modèle :

   $$\begin{aligned} P(O|\lambda) &= \sum_{Q \in S^T} P(O, Q|\lambda) \\ &= \sum_{Q \in S^T} P(O|Q, \lambda) P(Q|\lambda) \end{aligned} \tag{2.11}$$

   L'évaluation directe de cette probabilité est impossible en pratique, du fait de sa complexité calculatoire : elle nécessite en effet de l'ordre de $2T \times N^T$ opérations. Une évaluation exacte de cette probabilité peut cependant être obtenue par l'algorithme *Forward-Backward*.

2. **Apprentissage ou adaptation du modèle.** Étant données n séquences indépendantes d'observations O^k, on désire trouver le modèle λ qui maximise $P(O|\lambda)$. Le but de l'apprentissage est donc de déterminer les paramètres (A, B, Π) qui maximisent le produit

   $$\prod_{k=1}^{n} P(O^k|\lambda)$$

   Le critère à optimiser est celui du *maximum de vraisemblance* (*Maximum Likelihood Estimation*). Les solutions utilisées s'appuient sur des optimisations locales telles que les techniques du *gradient*, ou l'algorithme de *Baum-Welch*. Ce dernier est basé sur le théorème de Baum qui garantit l'atteinte d'un maximum local de la fonction de vraisemblance par estimations itératives des paramètres A, B, Π. La solution obtenue est fortement dépendante du HMM initial.

   A partir d'un ensemble d'apprentissage, on peut également estimer empiriquement le modèle λ, ce que nous ferons par la suite.

3. **Chemin le plus probable, ou reconnaissance d'une séquence, ou décodage d'une séquence.** Étant donnée la séquence d'observations O et un modèle λ, il s'agit de déterminer la séquence Q d'états cachés ayant la plus forte probabilité d'avoir généré la séquence O d'observations. C'est ce problème en particulier que nous allons traiter dans le cadre de la structuration de vidéo.

   La reconnaissance peut être effectuée de deux façons différentes : soit, dans le cas d'un modèle par classe, par recherche du modèle discriminant (*Model Discriminant*), soit, dans le cas d'un seul modèle pour toutes les classes, par recherche du chemin optimal qui fournira la classe (*Path Discriminant*).

   Dans le premier cas, la reconnaissance se fait simplement par le calcul des probabilités d'émission de la séquence d'observations par les modèles. La séquence à reconnaître est affectée à la classe dont le modèle fournit la probabilité la plus importante :

$$\lambda^* = \arg\max_{\lambda \in \Lambda} P(O|\lambda) \tag{2.12}$$

   Dans le deuxième cas, la reconnaissance consiste à déterminer le chemin le plus vraisemblable pour cette observation, c'est-à-dire à trouver la meilleure suite d'états, appelée *suite d'états de Viterbi*, qui maximise la quantité $P(Q|O,\lambda)$. Ceci revient à trouver le meilleur chemin dans un graphe. La structure de ce graphe se prête aux techniques de la programmation dynamique. Une technique largement employée est *l'algorithme de Viterbi*. C'est cette deuxième solution que nous allons utiliser, car elle permet de réaliser simultanément la segmentation et la classification.

   Dans ce cas, il s'agit d'estimer $\hat{Q}$:

$$\begin{aligned} \hat{Q} &= \arg\max_{Q \in S^T} P(Q|O,\lambda) \\ &= \arg\max_{Q \in S^T} \frac{P(O|Q,\lambda)P(Q|\lambda)}{P(O|\lambda)} \\ &= \arg\max_{Q \in S^T} P(O|Q,\lambda)P(Q|\lambda) \end{aligned} \tag{2.13}$$

   Les observations étant supposées conditionnellement indépendantes, on a :

$$\begin{aligned} P(O|Q,\lambda) &= \prod_{t=1}^{T} P(o_t|q_t,\lambda) \\ &= b_{q_1}(o_1) \cdot b_{q_2}(o_2) ... \cdot b_{q_T}(o_T) \end{aligned} \tag{2.14}$$

   La probabilité d'une séquence d'états Q donnée s'écrit :

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T} \tag{2.15}$$

   Il en découle que $\hat{Q}$ peut s'écrire :

$$\hat{Q} = \arg\max_{q_1, q_2, ..., q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_1}(o_2) ... a_{q_{T-1} q_T} b_{q_T}(o_T) \tag{2.16}$$

2.2.4 Autres problèmes liés aux HMMs

Lors de l'utilisation des HMMs, un certain nombre de problèmes annexes surviennent.

- **L'insuffisance de données pour l'apprentissage** : pour que les paramètres du HMM convergent vers les vraies valeurs, le nombre d'observations devrait tendre vers l'infini. Dans la pratique, le nombre d'observations est toujours fini.

- **La meilleure initialisation du HMM** : la solution donnée par l'algorithme de Baum-Welch est dépendante du modèle initial. Elle risque de converger vers un maximum local si le modèle initial est mal choisi. Un bon choix des estimations initiales réduit le nombre d'itérations et produit un meilleur modèle. Mais, il n'existe pas de méthodes théoriques permettant de trouver les bonnes valeurs initiales.

- **Le choix de l'architecture du HMM la mieux adaptée aux données**. La détermination du nombre d'états d'un HMM n'est pas toujours chose aisée. Il n'existe pas de méthode théorique permettant de déterminer le nombre exact d'états pour l'optimisation du HMM. De plus, les états ne sont pas toujours liés à un phénomène physique observable. Deux démarches sont suivies pour résoudre ce problème :
  - on effectue des tests avec différents nombres d'états pour sélectionner la valeur "optimale" ;
  - le HMM est relié à une explication physique, et on cherche le nombre d'états d'après les connaissances *a priori* du domaine. C'est cette deuxième démarche que nous allons adopter.

Nous allons maintenant expliquer comment les modèles de Markov cachés sont appliqués pour segmenter une vidéo en unités logiques.

2.3 Modélisation des unités logiques du tennis

Avant d'exposer l'analyse formelle des documents étudiés et la méthode que nous utilisons, nous informons le lecteur que les règles élémentaires du tennis sont résumées dans l'annexe A.

2.3.1 Détermination des unités logiques

Les règles de production et les règles du tennis sont utilisées pour définir et décrire les unités logiques d'une vidéo. Les différents plans d'une vidéo sont identifiés selon leur contenu (vue globale du terrain, gros plan, publicité...). Pour le tennis, on peut distinguer quatre catégories de prises de vue : les vues globales, les plans moyens, les gros plans et les vues du public (Fig 2.2).

Dans le cas du tennis, les unités intéressantes sont les phases d'échanges entre les joueurs. Durant un échange, le point de vue sélectionné est celui capturant la vue globale du terrain, tandis qu'entre les échanges, des plans rapprochés sur les joueurs ou sur le public seront préférés. Pour cette raison, l'unité logique du tennis est souvent définie par la vue globale du terrain. Cette unité logique étant composée d'un seul plan, la vue globale du terrain est appelée *vue canonique*. La plupart des approches cherchent à détecter

(a) vue globale (b) plan moyen (c) gros plan (d) public

FIG. 2.2: Les quatre principales prises de vue dans une retransmission de tennis.

ces vues [25, 45, 30]. Il s'agit là d'une analyse simple d'une vidéo de tennis qui part de l'hypothèse qu'une vue globale est une phase d'échange entre les joueurs.

Nous avons choisi de définir plusieurs unités logiques afin de permettre : (i) d'identifier plus de segments, et (ii) une structuration *dense* de la vidéo. Aussi, nous avons pris en compte quelques observations supplémentaires :

- toutes les vues du terrain n'ont pas le même intérêt. D'abord, l'hypothèse selon laquelle une vue du terrain représente une phase d'échanges entre les joueurs est mise en défaut lorsqu'il n'y a pas d'échange. Ensuite, un échange peut être avorté, typiquement par un mauvais service ;
- les rediffusions soulignent le caractère important de l'échange qui les précède ;
- les changements de côté des joueurs et les publicités sont un indice du statut du jeu : ils indiquent la fin d'un jeu, sinon d'un set.

La segmentation en unités logiques a pour objectif d'éviter le recours à des attributs coûteux comme la segmentation d'objet et leur suivi. Elle ne se base que sur des attributs bas-niveau et l'analyse de l'entrelacement temporel des plans. Il en résulte une certaine limitation quant à la nature des segments identifiés. Par exemple, cela ne suffit pas à caractériser un ace ou un retour gagnant, pas plus qu'une montée au filet.

A partir de ces remarques, nous avons identifié quatres scènes caractéristiques d'une vidéo de tennis :

- *les échanges* : il s'agit de vues globales au terme desquelles un point est marqué ;
- *les premiers services ratés* : ils sont caractérisés par une vue globale d'une courte durée, mais ne concluent pas un point ;
- *les temps morts* ou *temps de repos* : de durée significative, ils apparaissent lorsque les joueurs changent de côté (tous les deux jeux en général) ;
- *les rediffusions* : elles montrent la dernière action menée suivant un autre point de vue, ou au ralenti. Elles sont notifiées au téléspectateur par l'insertion de transitions spéciales.

Ces quatre scènes composent nos unités logiques. Nous allons voir dans la suite comment nous les avons caractérisées à partir des points de vues et des informations temporelles.

On constate que les différents plans ne se réalisent pas indépendamment les uns des autres au cours du temps : ils obéissent à une syntaxe. Par contre, connaissant l'unité logique à un instant donné (par exemple premier service raté), la suite du scénario est indépendante du passé. C'est exactement ce que modélisent les modèles de Markov d'ordre 1. Puisque ces unités logiques ne sont pas directement observables, mais caractérisées par les attributs visuels, on utilisera plus précisément les modèles de Markov cachés. Nous pré-

sentons la mise en œuvre des HMMs pour la modélisation des unités logiques dans le paragraphe suivant.

2.3.2 Modélisation des unités logiques

Nous considérons ici les plans, et non les images, comme unité temporelle de base d'une vidéo. La segmentation en unités logiques consiste à regrouper les plans adjacents en *scènes* (appelées ici unités logiques). La syntaxe des unités logiques désigne l'arrangement temporel des plans, résultant des règles de production de la vidéo.

Les HMMs modélisent de la façon suivante cette syntaxe :

- un état du HMM représente un plan de la vidéo. Les états d'un HMM sont différenciés par leur contenu (le point de vue qu'ils représentent) et leur position les uns par rapport aux autres ;
- les observations associées à un état sont les attributs que l'on peut extraire automatiquement du plan ;
- les transitions entre les états représentent la succession des plans dans la vidéo, *i.e.* leur entrelacement temporel.

Un HMM par unité logique est défini. La spécification totale d'un HMM nécessite la spécification des paramètres du modèle (N le nombre d'états et M le nombre de symboles d'observation), la spécification des symboles d'observations, et la spécification des trois matrices stochastiques : A, B et Π.

L'étape fondamentale est le choix de l'architecture du HMM la mieux adaptée aux données. Dans notre cas, le HMM est expliqué par les règles de production. Les connaissances *a priori* du domaine sont exploitées pour déterminer le nombre d'états et la topologie des HMMs. L'architecture des différents HMMs rend donc compte explicitement des règles utilisées.

Nous allons décrire précisément dans les paragraphes suivants, comment les HMMs ont été spécifiés pour intégrer l'information *a priori*.

2.3.2.1 Topologie des HMMs

La spécification des HMMs s'appuie sur les règles suivantes :

Règles de production – Si toutes les vues du terrain ne sont pas forcément synonymes d'un échange entre les joueurs, un échange est nécessairement capturé par une vue globale du terrain ;
- Entre deux échanges, durant les phases de non-jeu pendant lesquelles les joueurs se préparent pour l'échange suivant, des gros plans sur les joueurs ou le public, ou des plans rapprochés sont diffusés ;
- Les rediffusions sont notifiées au téléspectateur par l'insertion de transitions spéciales.

Règles du tennis relatives au déroulement du jeu [84]
- Le jeu doit être continu depuis le premier service jusqu'à la fin de la partie ;
- Après une éventuelle première faute de service, la seconde balle doit être servie sans délai ;
- Aux changements de côtés, il ne doit pas se passer plus d'une minute trente entre la fin du dernier point et le moment où la balle suivante est servie ;

– A la fin de chaque set, les joueurs ont droit à un repos de deux minutes entre la fin du dernier point et le moment où la balle suivante est servie.

Ces informations sont intégrées aux HMMs représentant les unités logiques de la façon suivante (Fig. 2.3) :

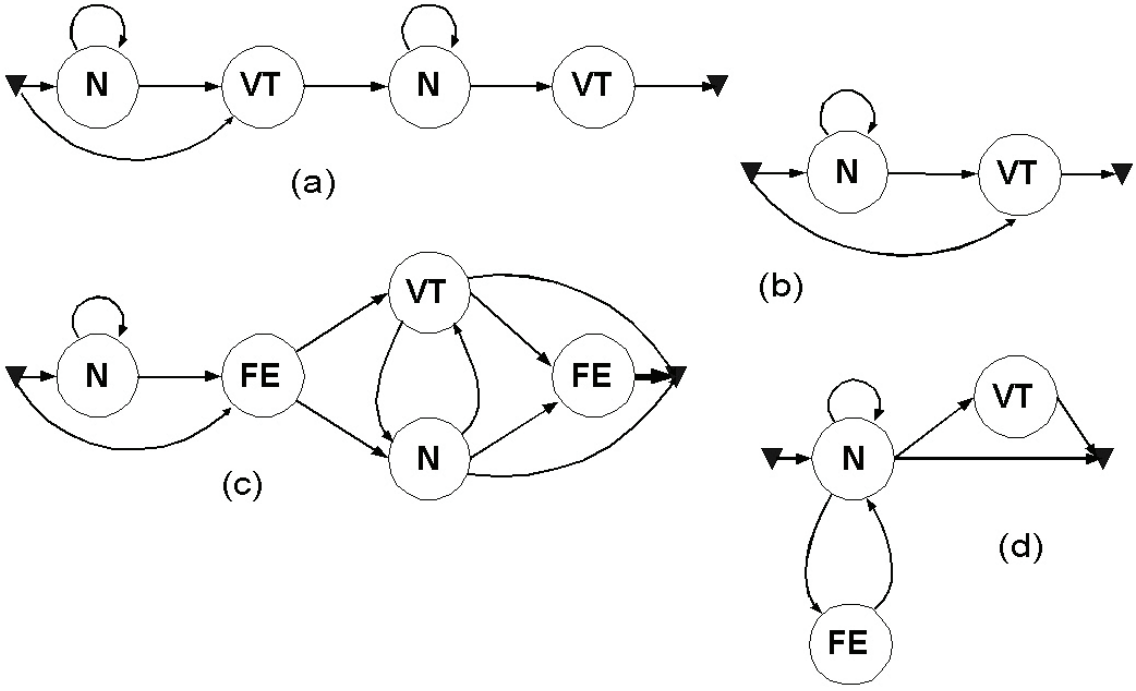


FIG. 2.3: Modèles de Markov cachés des unités logiques du tennis. (a) premier service manqué (b) échange (c) rediffusion (d) temps mort. VT désigne les vues globales du terrain, FE les fondus enchaînés et N les autres vues.

Les échanges

Ils sont caractérisés par une vue globale du terrain de durée variable, précédée de plans non relatifs à une phase de jeu pendant lesquels les joueurs se préparent.

Les premiers services manqués

Les premiers services ratés sont des vues globales du terrain caractérisés par une courte durée. En cela, rien ne les distingue d'un service gagnant (*ace*), excepté leur contexte temporel. En effet, un joueur venant de rater son premier service est déjà en place pour le deuxième. Et comme l'indique la règle ci-dessus, la seconde balle doit être servie sans délai. Ainsi un premier service raté est caractérisé par le scénario suivant : une vue globale du terrain de courte durée, suivie de gros plans ou de plans rapprochés de courtes durées également, immédiatement suivi d'une autre vue globale du terrain capturant le deuxième service.

Lors d'un service gagnant, le point est marqué, et les joueurs doivent se replacer sur le terrain pour le prochain engagement. Un service gagnant est donc différencié d'un premier service manqué par l'intervalle de temps qui le sépare de la prochaine vue du terrain. Pour cette raison la modélisation temporelle d'un premier service raté est dépendante de l'échange qui la suit.

Les rediffusions
Elles montrent la dernière action menée suivant un autre point de vue, ou au ralenti. Elles se distinguent par la présence d'au moins une transition spéciale (ici nous considérons les fondus enchaînés).

Les temps morts
Nous entendons par "temps morts", non pas les phases de non-jeu qui séparent deux échanges (et qui sont intégrées dans les HMMs précédents), mais les temps de repos associés aux changements de côté des joueurs. Les règles citées plus haut indiquent que ces temps de non-jeu spécifiques sont d'une durée significativement supérieure aux temps des phases de non-jeu correspondant à la préparation des joueurs entre deux échanges. Les temps de repos associés aux changements de côté constituent également les moments idéaux pendant lesquels les producteurs vont pouvoir insérer de la publicité.

2.3.2.2 Les observations associées aux états

Les observations associées aux états sont les attributs des plans de la vidéo. Les caractéristiques utilisées pour la conception des HMMs ci-dessus sont essentiellement le type de la prise de vue (vue globale du terrain, publicité, gros plans, plans rapprochés) et la durée du plan.

Les observations associées à chaque état sont alors :

1. un indice v caractérisant le type ou la classe de la prise de vue. Cet attribut est déterministe ou non, selon qu'une classification des plans soit effectuée préalablement ou non. Nous avons considéré différents cas dans la suite. Dans le schéma de la figure 2.3, seules deux classes de plans sont utilisées : les vues globales du terrain et toutes les autres vues qui ne sont pas différenciées ;
2. un indice l représentant la durée du plan en nombre d'images.

Les indices v et l sont extraits automatiquement pour chaque plan de la vidéo.

2.3.2.3 Probabilités d'observations et probabilités de transitions

La probabilité qu'une observation $o_t = \{v_t, l_t\}$ soit générée par un état s_j s'écrit comme la probabilité jointe des événements v_t et l_t sachant s_j. En supposant l'indépendance conditionnelle de v_t et l_t sachant l'état, il suit :

$$b_j(o_t) = P(o_t = \{v_t, l_t\}|s_j) = P(v_t|s_j) \cdot P(l_t|s_j) \tag{2.17}$$

Les probabilités $P(v_t|s_j)$ et $P(l_t|s_j)$ sont calculées à partir des lois de probabilités P_v et P_l qui sont estimées par apprentissage.

Les probabilités de transitions entre les états sont également estimées par apprentissage. En fait, la figure 2.3 illustre par souci de simplification les transitions les plus probables. Cependant, les modèles obtenus par l'apprentissage sont ergodiques, c'est-à-dire que tous les états sont interconnectés. Il n'y a pas d'information *a priori* explicite sur les transitions entre états.

Dans le paragraphe suivant, nous expliquons comment les indices l et v sont extraits automatiquement de chaque plan. La longueur l de chaque plan résulte directement de la segmentation en plans de la vidéo. L'indice v, relatif au type de la prise de vue, correspond à une mesure de similarité visuelle.

2.4 Caractérisation des plans

Comme il a été annoncé lors de la présentation du système (2.1.3), la première étape consiste à segmenter la vidéo en plans. Nous nous intéressons à deux types de transitions : les coupures et les fondus enchaînés.

Dans le paragraphe suivant, nous décrivons les méthodes utilisées pour extraire les attributs du plan que nous avons choisis : *l*a durée du plan, les fondus enchaînés, mais aussi *l*'activité du plan. Cette dernière n'est pas *stricto sensu* une observation, mais sera utilisée pour caractériser visuellement le plan.

2.4.1 Segmentation temporelle de la vidéo

Pour la segmentation temporelle et le calcul de l'activité, nous exploitons les propriétés du codage MPEG. La compression des séquences d'images est réalisée par l'application simultanée sur des blocs d'images d'une Transformée Cosinus Discrète (DCT), et d'une prédiction de mouvement. Trois types d'informations en sont déduites : les coefficients DC issus des blocs DCT, les vecteurs de mouvement et une classification du mouvement dominant en terme de zoom, translation, absence de mouvement ou inconnu.

2.4.1.1 Longueur du plan

Détection des coupures

Le procédé de détection des coupures est une simple méthode de différence d'histogrammes entre deux images successives. Un histogramme de luminance est calculé à partir de l'ensemble des coefficients DC de la transformée DCT. Les histogrammes sont lissés par un filtre passe-bas d'ordre 5. Soit H_1 et H_2 deux histogrammes, on définit la distance D_{trans} par :

$$D_{trans}(H_1, H_2) = \frac{\sum_i \left[\min\ \left(|H_2[i] - H_1[i-1]|, |H_2[i] - H_1[i]|, |H_2[i] - H_1[i+1]|\right)\right]}{\sum_i H_1[i]}$$

Une transition est détectée lorsque la distance entre deux images consécutives dépasse un seuil Th_{cut} prédéfini.

On utilise $Th_{cut} = 0.16$. La table 2.1 indique les résultats de la détection des coupures sur quelques unes des séquences utilisées. On définit :

$$\begin{aligned} \text{précision} &= 100 \cdot \frac{\#\text{correct}}{\#\text{correct} + \#\text{faux}} \\ \text{rappel} &= 100 \cdot \frac{\#\text{correct}}{\#\text{correct} + \#\text{manqué}} \end{aligned} \tag{2.18}$$

où #correct est le nombre de coupures correctement classifiées, #faux est le nombre de fausses classifications et #manqué est le nombre de coupures manquées. Cette méthode simple fournit de très bons résultats. Les fausses détections sont généralement dues à un phénomène d'occultation pendant la prise de vue (passage d'une personne, d'un objet devant le champ de la caméra).

Vidéos	Coupures					
	total	correct	manquées	fausses	*précision*	*rappel*
RGO2_set2	253	248	5	21	0.92	0.98
RGO2_set3	256	249	7	9	0.96	0.97
DavisCup_set1	470	457	13	11	0.97	0.97
DavisCup_set2	387	377	10	12	0.97	0.97
DavisCup_set3	390	381	9	9	0.97	0.97
RL01_set1	376	371	5	7	0.97	0.98
RL01_set2	437	432	5	10	0.97	0.98
US_Open_05	293	289	7	0	1	0.98
US_Open_09	314	304	10	2	0.99	0.96

TAB. 2.1: Evaluation de la segmentation temporelle - détection des coupures.

Détection des fondus enchaînés

La détection des fondus enchaînés est basée sur une méthode de double seuillage. Elle compare les histogrammes non pas d'images successives H_t, H_{t+1}, mais d'images distantes H_t, H_{t+k} [85]. Tout d'abord, un premier seuillage sélectionne une image de référence H_b comme candidate du début d'une transition potentielle, si elle vérifie : $D_{trans}(H_{b-1}, H_b) > Th_b$.

Ceci fait, toutes les images suivantes H_{b+k} sont comparées à H_b. Un fondu enchaîné est détecté lorsque $D_{trans}(H_{b+k}, H_b) > Th_e$. La fin de la transition est alors décidée lorsque $D_{trans}(H_{t+1}, H_t)$ est de nouveau inférieur à Th_b (Fig. 2.4).

Une transition en cours est annulée dans les deux cas suivants :

- lorsqu'une coupure, un zoom ou une translation est détectée. Cela évite que les mouvements de caméra, en entraînant une modification progressive du contenu du plan, ne déclenchent une suite de fausses détections ;
- si la différence $D_{trans}(H_{t+1}, H_t)$ entre deux images consécutives est inférieure à Th_b. On considére alors que l'image H_{t+1} n'est pas une image de transition. Comme il peut aussi s'agir d'une exception, on autorise une certaine tolérance sur le nombre d'images n'appartenant pas à la transition, par le biais d'un seuil Th_{res} sur le nombre d'exceptions.

La table 2.2 indique les résultats de la détection des fondus enchaînés sur quelques unes des séquences utilisées. Comme pour tout système basé sur un seuillage, l'inconvénient de cette méthode est l'ajustement des seuils, afin de trouver le bon compromis entre fausses alarmes et détections manquées. Les seuils Th_b, Th_e, et Th_{res} ont été fixés de façon à minimiser les fausses alarmes, cependant les résultats varient d'une séquence à l'autre. La majorité des fausses alarmes provient de l'insertion brusque ou progressive d'incrustations (du score notamment) dans une partie de l'image. Adapter précisément les seuils à chaque séquence est une tâche laborieuse et illusoire, et il nous semble plus réaliste de prendre en compte les imperfections de la segmentation dans la suite du processus de structuration.

En conclusion, notons que les meilleurs taux de détection atteignent des valeurs de 95-98%, pour des taux de fausses alarmes souvent très élevés (150-500%). Enfin, malgré le grand nombre de techniques disponibles actuellement, la détection des transitions, en particulier progressives, reste problématique. Elle se heurte à des problèmes de sensibi-

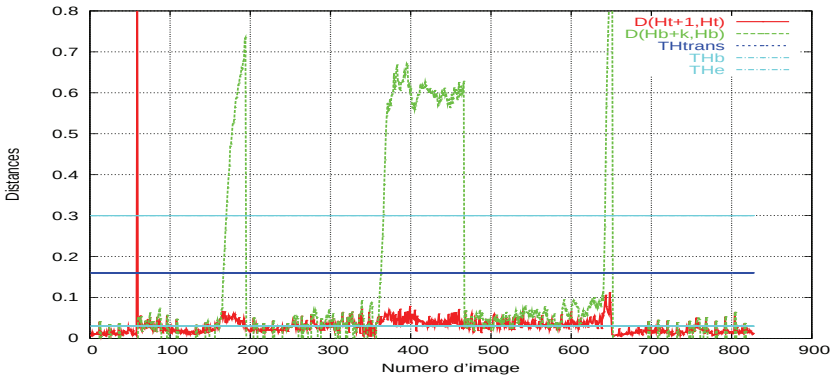


FIG. 2.4: Diagramme de détection des fondus enchaînés - détection d'1 coupure et de 3 transitions progressives

lité aux mouvements de caméra et d'objets présents dans la scène, ou aux changements d'illuminations (comme les flashes).

Vidéos	Transitions Progressives					
	total	correct	manquées	fausses	*précision*	*rappel*
RGO2_set2	54	27	27	3	0.9	0.5
RGO2_set3	30	19	11	2	0.9	0.63
DavisCup_set1	71	51	20	12	0.8	0.71
DavisCup_set2	43	26	17	15	0.63	0.6
DavisCup_set3	39	28	11	21	0.57	0.71
RL01_set1	42	34	8	12	0.74	0.8
RL01_set2	33	20	13	14	0.58	0.6
US_Open_05	57	39	18	14	0.73	0.68
US_Open_09	36	23	13	14	0.62	0.64

TAB. 2.2: Evaluation de la segmentation temporelle - détection des transitions progressives avec $Th_b = 0.03$, $Th_e = 0.3$ et $Th_{res} = 2$.

2.4.1.2 Activité

Nous avons choisi l'activité comme mesure globale du mouvement dans le plan. L'activité est mesurée directement lors du décodage de la séquence.

Les vecteurs de mouvement encodés dans les flux MPEG représentent les déplacements entre deux blocs de pixels, l'un appartenant à l'image courante et l'autre à l'image de référence. Ces vecteurs servent à estimer le mouvement dominant apparent dans la vidéo. Cependant, l'utilisation des vecteurs de mouvement MPEG pose quelques problèmes, dans la mesure où ils n'ont pas nécessairement de signification physique. Pour cette raison, les vecteurs aberrants au sens du mouvement dominant sont supprimés par une régression linéaire robuste.

L'activité est définie comme la moyenne sur l'image de l'amplitude des vecteurs de

mouvement MPEG non aberrants associés à l'image :

$$a = \sum_{(u,v)\in\Omega} \frac{u^2 + v^2}{|\Omega|} \tag{2.19}$$

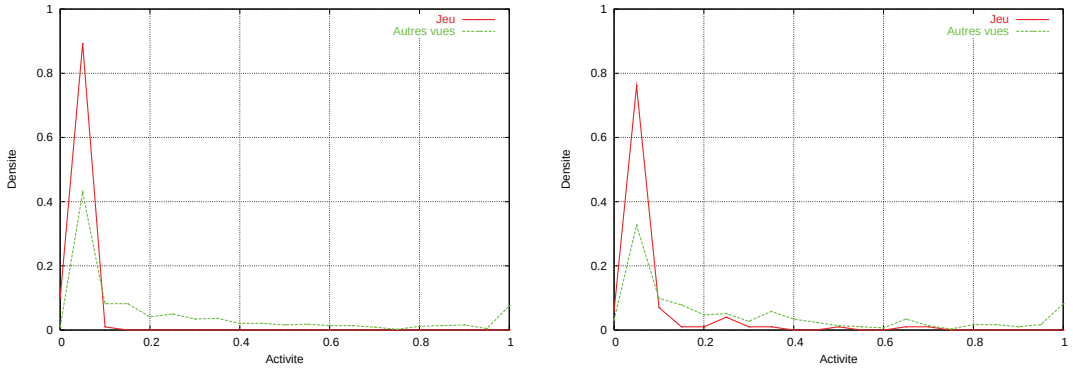
où (u, v) sont les composantes horizontales et verticales du vecteur de mouvement associé au pixel (x, y) de l'image, Ω est l'ensemble des vecteurs valides et $|\Omega|$ son cardinal.

La distribution des valeurs prises par l'activité est similaire en fonction du type de plans. Cependant, elles dépendent du type de vidéo, et notamment de l'encodage. De plus, il existe quelques valeurs isolées très importantes.

L'activité est donc seuillée et normalisée par un facteur de normalisation Norm dépendant de la séquence et calculé automatiquement :

$$\text{Norm} = \bar{a} + 1.96 \cdot \sigma_a \tag{2.20}$$

où $\bar{a}$ est l'activité moyenne sur la séquence traitée et σ_a son écart-type.



(a) Données extraites d'un set de la Coupe Davis

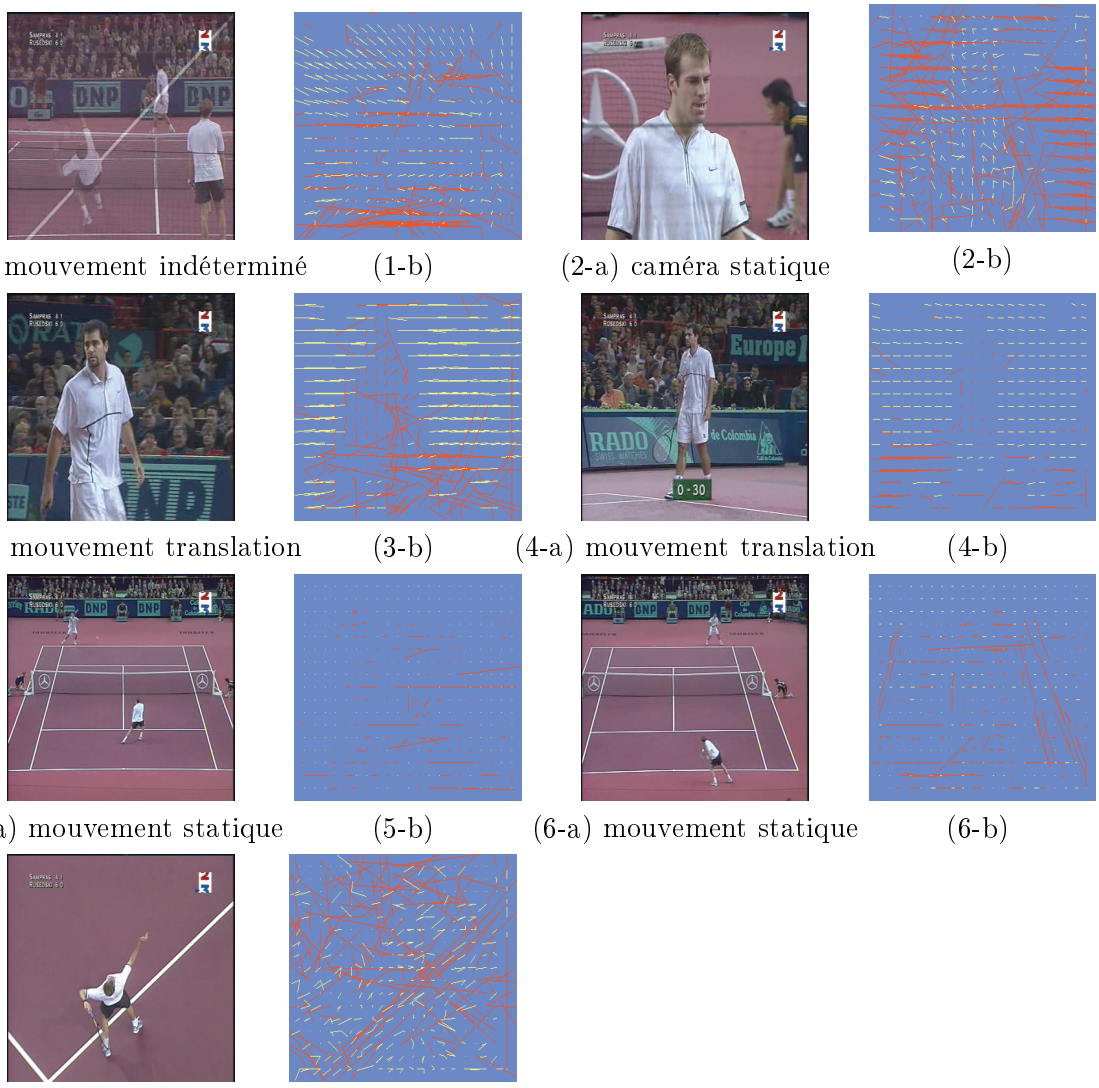
(b) Données extraites d'un set de l'Open de Paris

FIG. 2.5: Répartition de l'activité normalisée selon le type de plan pour 2 vidéos de sources différentes.

Au tennis, les vues globales du terrain doivent capturer la plus grande partie du court, tandis que les gros plans suivent généralement un des joueurs. Ainsi les vues globales présentent un faible mouvement de caméra, tandis que les autres plans sont plus susceptibles de présenter d'importantes translations de la caméra. Cette caractérisation du contenu n'est cependant pas suffisamment fiable comme observation en entrée des modèles de Markov. Sur un gros plan, le mouvement dominant est faible si les vecteurs valides sont ceux du visage, ou au contraire important si le mouvement dominant est celui de l'arrière-plan. Nous utilisons donc l'activité simplement comme un indice supplémentaire dans le calcul de la similarité visuelle d'un plan tel qu'elle est définie dans le paragraphe suivant.

2.4.2 Similarité visuelle

Le contenu visuel du plan est représenté par son image-clé. Les caractéristiques visuelles sont utilisées pour identifier les vues globales parmi toutes les images-clé. Les vues globales



(1-a) mouvement indéterminé (1-b) (2-a) caméra statique (2-b)

(3-a) mouvement translation (3-b) (4-a) mouvement translation (4-b)

(5-a) mouvement statique (5-b) (6-a) mouvement statique (6-b)

(7-a) mouvement statique (7-b)

FIG. 2.6: Vecteurs mouvements MPEG : en rouge, les vecteurs aberrants, en jaune, les vecteurs valides conservés pour calculer l'activité.

doivent capturer à chaque instant la plus grande partie du court. Elles se caractérisent donc par un contenu couleur relativement homogène, puisqu'il s'agit de la couleur du terrain et de son environnement, à l'inverse des plans rapprochés et des plans du public.

Nous présentons ici l'attribut que nous avons utilisé pour décrire le contenu visuel du plan. Il s'agit d'une mesure de similarité permettant de caractériser les vues du terrain. Cette mesure s'appuie sur un descripteur image qui est **un vecteur de couleurs dominantes**.

Le processus d'extraction se divise en trois étapes. D'abord, les couleurs dominantes sont extraites pour chaque image-clé. Ensuite une image-clé K_{ref} représentative d'une vue globale est sélectionnée automatiquement, sans aucune hypothèse sur la couleur du terrain. Une fois que K_{ref} a été sélectionnée, la similarité visuelle entre chaque image-clé de la séquence et K_{ref} est calculée.

2.4.2.1 Extraction des couleurs dominantes

L'espace de représentation des couleurs est YCbCr qui est l'espace de représentation des signaux vidéo pour la diffusion notamment utilisé par le schéma de compression MPEG. L'espace YCbCr permet de s'affranchir des variations de la puissance lumineuse pour traiter l'information de teinte. Il s'appuie sur la luminance et sur deux canaux de chrominance (CbCr). L'importance de cet espace est aussi liée au fait que le système visuel humain perçoit le stimulus couleur plutôt en termes de luminance et chrominance qu'en terme d'attributs rouge, vert et bleu.

Nous avons choisi de représenter l'information couleur sous forme de couleurs dominantes pour les raisons suivantes :
- il s'agit d'une représentation compacte (comparée à un histogramme par exemple) ;
- on peut introduire une notion de distribution spatiale de la couleur ;
- elle s'applique particulièrement bien au problème d'identification des vues du terrain, dominées par la présence de la couleur du terrain.

L'attribut couleur d'une image-clé i possède donc deux composantes :
- un vecteur de couleurs dominantes de l'image-clé F_i ;
- sa cohérence spatiale C_i.

F_i est un vecteur de N_i couleurs dominantes représentées par leur coordonnées dans l'espace YCbCr et leur pourcentage p_i dans l'image. Les couleurs de l'image originale sont quantifiées en N_i valeurs par un algorithme k-means. Deux couleurs voisines sont fusionnées lorsque leur distance est inférieur à un seuil T_d. Cela assure que les N_i couleurs dominantes obtenues sont perceptuellement différentes. La distance entre deux vecteurs de couleurs dominantes F_1 et F_2 est mesurée par la distance quadratique simplifiée suivante :

$$d^2(F_1, F_2) = \sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2a_{1i,2j}p_{1i}p_{2j} \tag{2.21}$$

où $a_{k,l}$ est un coefficient de similarité entre les couleurs c_k et c_l :

$$a_{k,l} = \begin{cases} 1 - \frac{d_{k,l}}{d_{max}} & \text{si } d_{k,l} \leq T_d \\ 0 & \text{si } d_{k,l} > T_d \end{cases}$$

T_d est la distance maximum pour laquelle on considère que deux couleurs sont similaires, $d_{max} = \alpha\, T_d$ et $d_{k,l}$ est la distance euclidienne entre deux couleurs c_k et c_l.

Afin de prendre en compte la configuration spatiale des pixels de couleurs similaire, une mesure de cohérence CO_k est calculée pour chaque couleur c_k d'un vecteur de couleurs dominantes. Un pixel de couleur c_k est considéré comme cohérent si tous les pixels de son voisinage ont la même couleur. CO_k est donc défini par :

$$CO_k = \frac{\text{Nombre de pixels cohérents de couleur } c_k}{\text{Nombre total de pixels de couleur } c_k} \tag{2.22}$$

Par suite, la mesure de confiance C_i pour un vecteur de couleurs dominantes F_i est :

$$C_i = \sum_{k=1}^{N_i} CO_k \times p_k \tag{2.23}$$

Dans notre implémentation, les vecteurs de couleurs dominantes sont de taille N=4 (Fig. 2.7), taille minimum nécessaire pour identifier le terrain.

(a) vue globale (b) plan moyen (c) gros plan (d) public

FIG. 2.7: Extraction de 4 couleurs dominantes.

2.4.2.2 Sélection d'une image-clé de référence

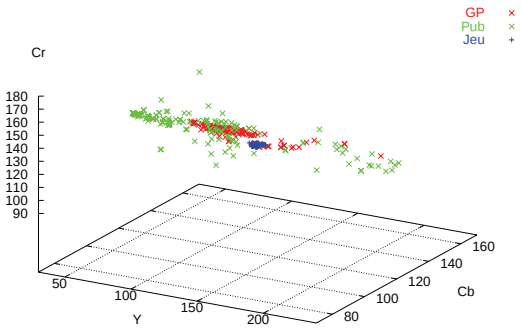
Les vues globales ne représentent en moyenne que 20% de l'ensemble total des vues, notamment à cause de la publicité aux plans courts produisant de nombreuses images-clés. Néanmoins, les vues globales présentent deux caractéristiques exploitées dans la suite :

- leur couleur dominante est celle du terrain, et couvre plus de 50% de l'image ;
- elles forment un groupe compact au sens de la distance définie en (2.21) (voir figure 2.8).

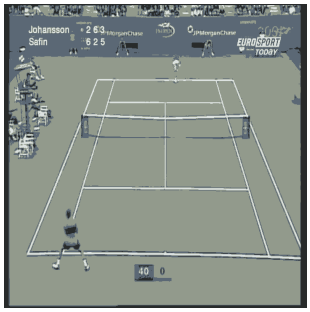
Afin de trouver une image-clé K_{ref} représentative d'une vue globale, une première sélection est effectuée sur les pourcentages de couleurs dominantes de chaque image-clé. Les images-clés dont le plus important pourcentage de couleur dominante est inférieur à 50% sont éliminées. Dans l'ensemble Ω des images-clés candidates restantes, celles représentatives d'une vue globale sont majoritaires (voir figure 2.9).

L'image clé K_{ref} la plus représentative d'une vue globale au sens de la similarité visuelle est l'image-clé la plus proche du centre de gravité de l'ensemble I des images-clés de Ω représentatives d'une vue globale, autrement dit l'image qui vérifie :

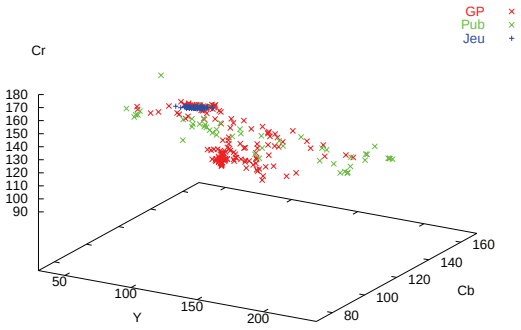
$$K_{\text{ref}} = \arg\min_{j \in I} \sum_{i \in I} d^2(F_i, F_j)$$



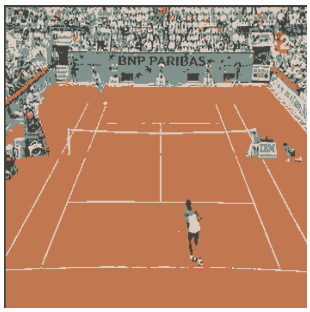
(a) US Open



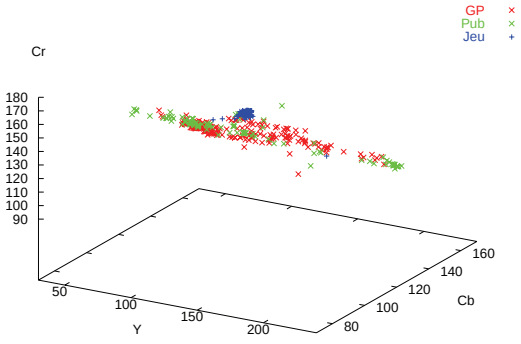
(b) Vue du terrain



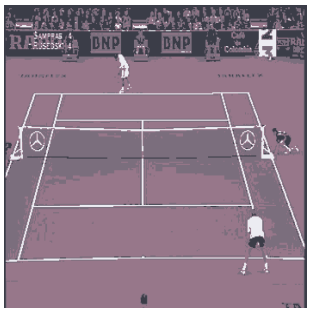
(a) Roland Garros



(b) Vue du terrain

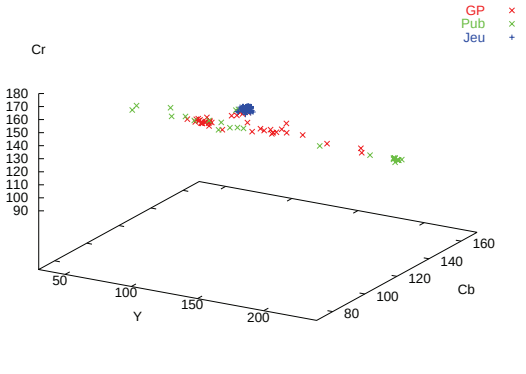


(a) Open Paris

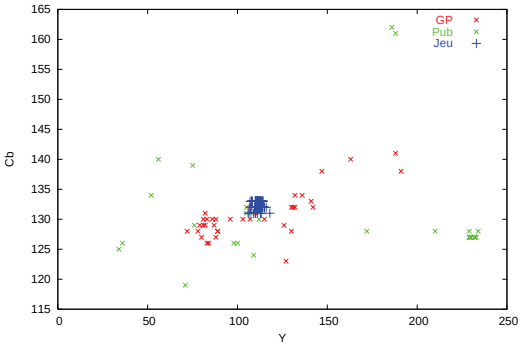


(b) Vue du terrain

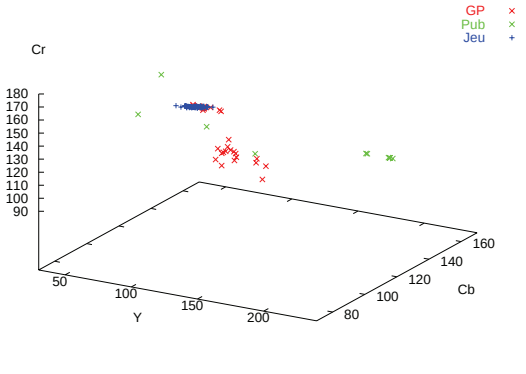
FIG. 2.8: Répartition de la couleur dominante en fonction du type de plan.



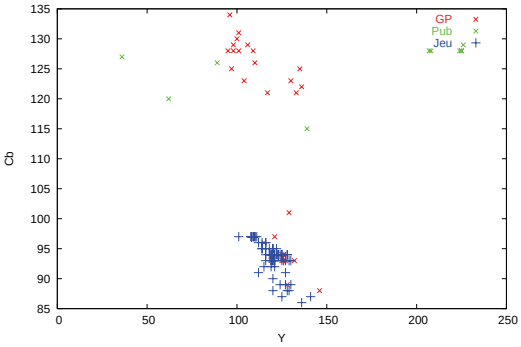
(a) Représentation dans l'espace YCbCr



(b) Projection sur l'espace YCb



(a) Représentation dans l'espace YCbCr



(b) Projection sur l'espace YCb

FIG. 2.9: Répartition de la couleur dominante dans le sous-ensemble des plans dont le pourcentage de la couleur dominante est supérieure à 50%- en haut sur des données issues de l'Open de Paris, en bas de Roland Garros.

Cela revient à appliquer la méthode des moindres-carrés classique dans l'ensemble I. Cependant, on cherche K_{ref} dans l'ensemble Ω, et la moyenne quadratique est très sensible aux points aberrants (ici, l'ensemble des images appartenant à $\{\Omega - I\}$). Nous utilisons la méthode robuste des moindre carrés médians car les images-clés de I sont majoritaires dans Ω.

Les moindres carrés médians sont appliqués une première fois afin d'isoler les images aberrantes. Ils sont appliqués une seconde fois pour déterminer K_{ref} parmi les images restantes.

Le processus de sélection est le suivant :

1. tirage aléatoire de p images-clés $K_j \in \Omega$;
2. sélection de l'image K_m, telle que :

$$F_m = \arg \min_{j \in [1..p]} \underset{i \in \Omega}{\text{médiane}} \quad d^2(F_i, F_j)$$

3. formation du sous-groupe ε des images de Ω vérifiant :

$$d^2(F_i, F_m) < \underset{i \in \Omega}{\text{médiane}} \quad d^2(F_i, F_m)$$

   On a $\varepsilon \subseteq I$.

4. minimisation de la distance médiane dans l'ensemble ε et sélection de K_{ref} telle que :

$$F_{\text{ref}} = \arg \min_{j \in \varepsilon} \underset{i \in \varepsilon}{\text{médiane}} \quad d^2(F_i, F_j)$$

Dans la première étape, on procède à un tirage aléatoire sans remise afin de s'épargner le calcul exhaustif de toutes les distances à toutes les images de Ω. Il a été montré dans [86] que p peut être déterminé en fonction :

- du nombre de paramètres q à estimer ;
- du taux de contamination τ relatif au pourcentage de données aberrantes (τ est fixé *a priori*) ;
- de la probabilité P d'avoir au moins un tirage sans données aberrantes par la relation :

$$p = \frac{\log(1 - P)}{\log(1 - (1 - \tau)^q)} \tag{2.24}$$

Dans notre cas, $q = 1$ et on suppose le taux de contamination $\tau = 50\%$. Le nombre de tirages à réaliser est alors de 10 tirages pour une probabilité $P = 99.9\%$ que l'un de ces tirages soit une image-clé représentant une vue globale.

Par ce procédé, la première image sélectionnée K_m est une image de l'ensemble I. Cependant les étapes 3 et 4 assurent l'unicité de K_{ref} quelque soit le tirage. Autrement dit, plusieurs tirages sur une même séquence produisent différentes K_m, mais suite aux étapes 3 et 4, K_{ref} est la même.

2.4.2.3 Calcul de la similarité visuelle

Finalement, la mesure de similarité visuelle v(K_1, K_2) entre deux images-clé K_1 et K_2 est définie comme une fonction pondérée de la cohérence spatiale, de la distance entre les vecteurs de couleurs dominantes et de l'activité :

$$v(K_1, K_2) = w_1|C_1 - C_2| + w_2 d(F_1, F_2) + w_3 \frac{\|A_1 - A_2\|}{A_{max}} \quad (2.25)$$

où w_1, w_2, et w_3 sont des coefficients de pondération. La similarité visuelle $v(K_t, K_{\text{ref}})$, notée v_t dans la suite, est calculée entre chaque image-clé K_t et K_{ref}.

2.5 Modélisation des unités logiques du baseball

Nous allons illustrer dans cette partie comment la modélisation de l'entrelacement temporel des plans est appliquée à la construction d'unités logiques du baseball. Nous informons le lecteur que les règles du baseball sont résumées dans l'annexe B

2.5.1 Unités logiques du baseball

Au baseball, les cycles de bases sont le lancer de la balle suivi, éventuellement, de la frappe du batteur. Comme pour le tennis, les lancers et les frappes sont caractérisés par des points de vue spécifiques, et les lancers, comme les frappes peuvent être ratés.

Les unités logiques sont :

- *les lancers* : il s'agit des vues du lancer. Ils caractérisent un lancé qui n'est pas frappé par le batteur, soit que le batteur n'ait pas réussi, soit que le lancer soit faux ;
- *les lancers frappés* par le batteur. Selon que la balle soit rattrapée au vol ou non, le jeu continue sur le champ intérieur où les joueurs courent de base en base ;
- *les rediffusions*, comme pour le tennis, montrent la dernière action menée suivant un autre point de vue, ou au ralenti ;
- *les temps morts* : de durée significative, ils apparaissent lorsque le batteur change ou lorsque l'équipe attaquante passe en défense.

2.5.2 Modélisation des unités logiques

Les différents points de vues sont plus nombreux parce que la superficie du terrain est supérieure à celle du tennis, et le nombre de joueurs plus important. Si seule la vue du terrain est une vue active au tennis, réellement porteuse d'une information sur le jeu, il y a lieu de distinguer plusieurs vues actives au baseball :

- *la vue du lancer* ***L*** est une vue canonique du baseball ;
- *les vues du grand champ* ***GC***, lorsqu'elles surviennent juste après une vue du lancer, indiquent que la balle a été frappée par le batteur et que la caméra suit son mouvement ;
- *les vues du champ intérieur* ***CI***, de la même façon que les vues du grand champ, indiquent que la balle a été frappée lorsqu'elles surviennent juste après une vue du lancer. Après quelques plans, elles capturent la course des coureurs ;
- *les plans rapprochés* ***PR*** suivent soit un coureur, soit l'action sur une base ;
- *les gros plans* ***GP*** indiquent en général la fin d'une action.

Ces plans ont donc une signification différente selon le contexte dans lequel ils apparaissent. Ainsi la séquence :

$$L \rightarrow GP \rightarrow P \rightarrow GC$$

où P indique une vue du public, n'a pas la même signification que la séquence :

$$L \rightarrow GC \rightarrow GP \rightarrow P$$

bien que les plans en jeu soient les mêmes. La première représente un lancer non frappé, tandis que la deuxième représente un lancer frappé par le batteur et probablement attrapé de volée par la défense.

Le nombre supérieur de vues actives par rapport au tennis implique la nécessité de mettre en œuvre plus d'attributs bas-niveau pour les distinguer et une modélisation plus complexe des unités logiques. La figure 2.10 illustre une modélisation possible, quoique simplifiée, des unités logiques définies précédemment : un lancer, un lancer frappé et joué, les rediffusions et les temps morts. Concernant le lancer frappé et joué, il est envisageable de distinguer différents modèles selon que la balle est rattrapée au vol ou non.

2.6 Résultats Expérimentaux

2.6.1 Protocole expérimental

2.6.1.1 Présentation des données de test

Nos données de tests se composent de 12 séquences vidéos de tennis pour une durée totale de $7^h14'30"$ (Tab. 2.3). Ces vidéos sont extraites de 4 tournois différents, mais chaque séquence d'un même tournoi n'appartient pas nécessairement au même match. L'ensemble des séquences constitue donc 4 familles avec des surfaces de terrain et des styles de production différents. Le nombre de plans n'est pas proportionnel à la durée de la séquence. Il dépend du style de production et du nombre d'insertions de publicités. Le signal audio n'est pas disponible pour toutes les séquences.

Séquences vidéo	Durée	Nombre de plans	Audio	Tournoi
US_Open_05	46' 54"	490		US Open
US_Open_09	45' 01"	485		US Open
RGO2_set2	25' 40"	300		Roland Garros
RGO2_set3	20' 28"	282		Roland Garros
RG01_set1	30' 47"	522	√	Roland Garros
RG01_set2	32' 52"	540	√	Roland Garros
DavisCup_set1	42' 22"	606	√	Coupe Davis
DavisCup_set2	41' 18"	485	√	Coupe Davis
DavisCup_set3	39' 07"	502	√	Coupe Davis
OpenParis_set1	34' 07"	439	√	Open de Paris
OpenParis_set2	47' 47"	677	√	Open de Paris
OpenParis_set3	28' 07"	400	√	Open de Paris

TAB. 2.3: Séquences vidéos utilisées.

Cinq séquences sont utilisées pour l'apprentissage et les 7 autres sont réservées à l'évaluation. Parmi les 4 familles de séquences, l'une est totalement exclue de l'ensemble d'apprentissage, afin de pouvoir estimer les capacités de généralisation du modèle.

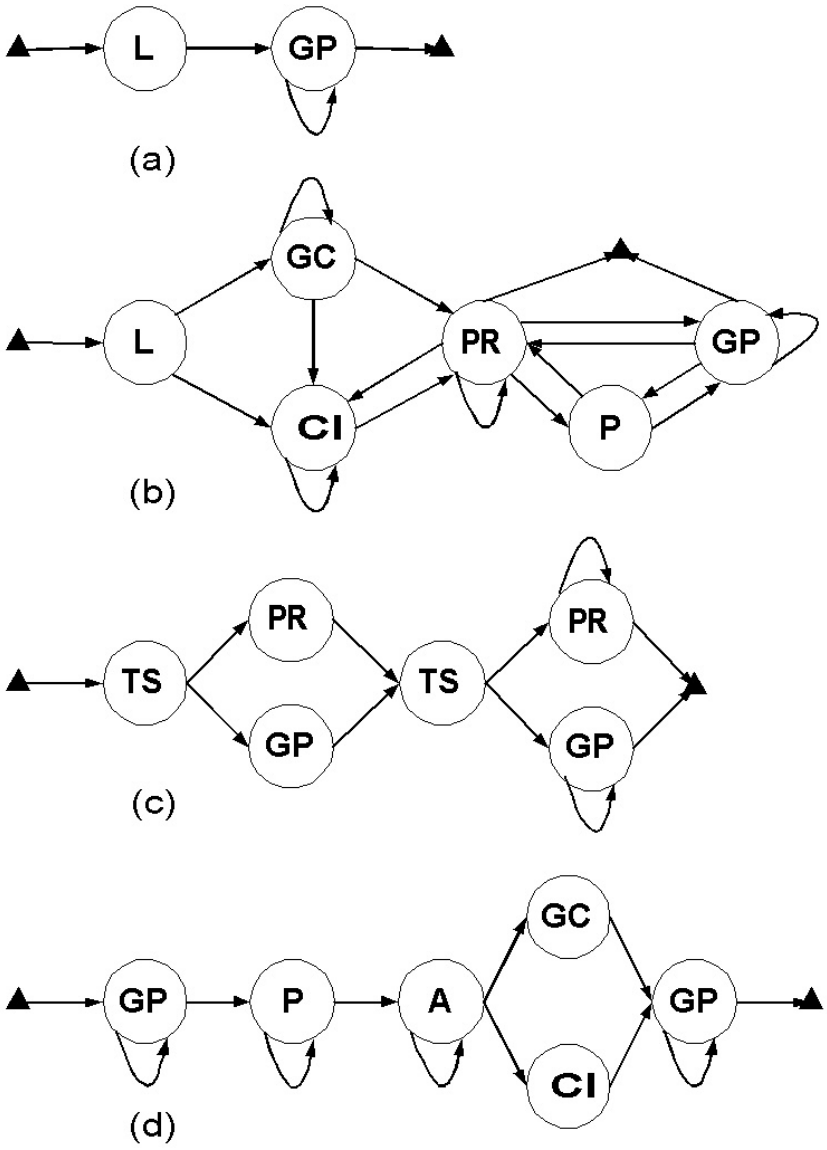


FIG. 2.10: Modèles de Markov cachés des unités logiques du baseball. (a) lancer (b) frappe (c) rediffusion (d) temps mort. L désigne les vues du lancers, GP les gros plans, PR les plans rapprochés, GC les vues du grand champ, CI les vues du champ intérieur, P les vues du public, TS les transitions spéciales et A les autres vues.

2.6.1.2 Apprentissage et décodage

La séquence d'observations est la suite des attributs extraits pour chaque plan. Soit une vidéo composée de T plans, on a :

$$O = (o_1, o_2, ...o_T) = (\{v_1, l_1\}, \{v_2, l_2\}, ..., \{v_T, l_T\})$$

où v_i est la similarité visuelle définie par l'équation 2.25 entre l'image-clé du plan i et l'image-clé de référence K_{ref} et l_i est la longueur du plan i.

Toutes les séquences ont été annotées manuellement afin d'estimer les performances de l'algorithme mis en place. L'apprentissage est effectué en alignant les états sur les données annotées afin d'estimer les paramètres du HMM. L'estimation des distributions de probabilités est réalisée par une méthode empirique de comptage des occurences :

$$\begin{aligned}
\hat{\pi}_i &= \text{fréquence de l'état } s_i \text{ à l'instant } t = 1 \\
\hat{a}(i,j) &= \frac{\text{nombre de transitions de } s_i \text{ vers } s_j}{\text{nombre d'occurrences de l'état } s_i} \\
\hat{b}_j(k) &= \frac{\text{nombre d'observations du symbole } v_k \text{ dans l'état } s_j}{\text{nombre d'occurrences de l'état } s_j}
\end{aligned}$$

Afin de pallier au problème d'insuffisance de données pour l'apprentissage, les histogrammes des distributions de probabilités sont lissés.

Une fois l'apprentissage effectué, une nouvelle séquence d'observation $O = (o_1, o_2, \ldots, o_T)$ est présentée au HMM. Le décodage consiste à calculer la séquence d'états $Q = (q_1, q_2, \ldots, q_T)$ telle que la probabilité $Pr(Q, O|\lambda)$ soit maximale, *i.e.* à fournir la séquence d'états réalisant le plus probablement la séquence d'observations. Les états q_t appartiennent à l'ensemble des états définis précédemment en 2.3.2 dans la figure 2.3. Nous avons utilisé l'algorithme classique de Viterbi. Celui-ci permet de trouver la séquence d'états $\hat{Q}$ la plus vraisemblable :

$$\hat{Q} = \arg\max_q \ \ln p(q) \ + \ \sum_t \ln b_{q_t}(o_t) \tag{2.26}$$

La séquence d'état $\hat{Q}$ issue du décodage est alors comparée à la séquence d'états annotée manuellement.

2.6.1.3 Expérimentations

Nous proposons deux approches de segmentation en unités logiques. Dans la première, les plans sont préalablement classifiés en "vues du terrain"/"autres vues". La séquence des observations est alors la séquence des attributs :

$$O = (o_1, o_2, \ldots, o_T) = (\{c_1, l_1\}, \{c_2, l_2\}, \ldots, \{c_T, l_T\})$$

où c_i est la classe du plan, et l_i est la longueur du plan i. Dans ce cas, la loi de probabilité P_c associée à l'observation c_t est binaire :

$$p(c_t|s_i) = \begin{cases} 1-\epsilon & \text{si l'état } s_i \text{ correspond à la classe } c_t \\ \epsilon & \text{sinon} \end{cases} \tag{2.27}$$

Dans la seconde méthode, les plans sont caractérisés par la similarité visuelle. La classification en terme de "vues du terrain"/"autres vues" est réalisée simultanément à la segmentation en unités logiques.

Nous présentons d'abord les méthodes de classification des plans, puis la segmentation en unités logiques d'après les deux approches précédemment définies.

2.6.2 Classification des plans

Dans un cadre général, la classification des plans nécessite de résoudre au moins deux problèmes majeurs : il faut fixer *a priori* le nombre de classes, sachant que les classes sémantiques ne sont pas forcément celles trouvées par un algorithme de classification, et il faut identifier *a posteriori* ces classes.

Dans le cas du tennis, seules deux classes sont réellement pertinentes : la classe des vues du terrain et la classe des autres vues. Nous présentons ici quelques méthodes simples, directes et déterministes de classification des plans, ne nécessitant pas d'apprentissage. Nous exploiterons les attributs extraits dans le paragraphe précédent : couleurs dominantes et mesure de similarité.

2.6.2.1 Méthode 1 : Quantification du pourcentage de pixels du terrain

Une méthode souvent utilisée est la classification selon le taux de pixels représentant la couleur du terrain dans chaque image. Cela permet une classification en trois niveaux : vues globales du terrain, plans moyens et gros plans. Ces trois classes sont caractérisées sur les hypothèses suivantes :

- les vues globales possèdent une majorité de pixels du terrain ;
- les plans moyens possèdent des pixels du terrain, mais pas en nombre majoritaires ;
- les vues qui ne possèdent pas de pixels du terrain.

Il faut noter que les publicités ne sont pas pré-détectées, bien qu'il existe des méthodes le permettant [87, 88]. Elles sont incluses dans la troisième classe.

Cette méthode nécessite de connaître la couleur du terrain c_f. Nous pouvons l'estimer grâce à l'extraction de K_{ref} présentée précédemment. Cette couleur variant légèrement d'une image à l'autre, il faut fixer un seuil T_{col} permettant de déterminer si une couleur c est la couleur du terrain : $d(c, c_f) < T_{col}$ où d est la distance euclidienne entre les couleurs c et c_f. Un second seuil T_l sur le pourcentage de pixels ayant la couleur c_f détermine s'il s'agit d'une vue globale du terrain ou non.

Le résultat de la classification est présenté dans le tableau 2.4, pour $T_{col} = 0.02$ et $T_l = 50\%$.

2.6.2.2 Méthode 2 : Seuillage de la similarité visuelle

Une autre méthode utilise directement la mesure de similarité entre chaque image et une image représentative d'une classe. Cela nécessite de disposer d'une image représentative de chaque classe, ce qui n'est pas évident à obtenir de façon automatique. Nous utilisons à nouveau l'image K_{ref} . Nous nous contenterons donc d'identifier deux classes : celle des vues globales et les autres. Cette méthode nécessite de fixer un seuil T_l sur la mesure de similarité permettant d'identifier les images proches de l'image modèle.

Le résultat de la classification est présenté dans le tableau 2.5, pour $T_l = 0.45$.

Vidéos	Vues du terrain		Autres vues	
	précision	*rappel*	*précision*	*rappel*
US_Open_05	97	95	98	99
US_Open_09	97	94	98	99
RGO2_set2	82	65	89	95
RGO2_set3	91	67	90	97
RG01_set1	96	63	90	99
RG01_set2	93	84	96	98
DavisCup_set1	98	97	99	99
DavisCup_set2	100	96	99	100
DavisCup_set3	98	97	99	99
OpenParis_set1	92	91	97	97
OpenParis_set2	87	93	98	96
OpenParis_set3	90	81	95	97
moyenne	93	85	95	98

TAB. 2.4: Résultats de la classification des plans en classes "vue du terrain"/"autre vue" par quantification du pourcentage de pixels du terrain.

Vidéos	Vues du terrain		Autres vues	
	précision	*rappel*	*précision*	*rappel*
US_Open_05	91	98	99	97
US_Open_09	96	98	99	99
RGO2_set2	71	76	92	90
RGO2_set3	64	72	90	86
RG01_set1	95	62	90	99
RG01_set2	90	77	94	98
DavisCup_set1	87	100	100	96
DavisCup_set2	95	97	99	98
DavisCup_set3	93	98	99	98
OpenParis_set1	84	100	100	94
OpenParis_set2	92	100	100	97
OpenParis_set3	81	97	99	94
moyenne	87	90	97	96

TAB. 2.5: Résultats de la classification des plans en classes "vue du terrain"/"autre vue" par seuillage de la similarité visuelle.

Les deux méthodes présentent des résultats équivalents. La méthode 2 est préférable car elle est moins complexe et ne nécessite de fixer qu'un seul seuil.

2.6.3 Segmentation en unités logiques

2.6.3.1 Avec une classification préalable des plans

Les séquences d'observations sont les durées et les classes (jeu ou non-jeu) des plans. Pour tester la validité du modèle, la segmentation est d'abord réalisée sur les données annotées manuellement. En utilisant uniquement les attributs de classe et de durée, le taux de classification correcte en unités logiques varie de 54% à 73%. Étant donné que les observations utilisées sont fiables, les mauvaises classifications sont uniquement dues aux transitions entre les états. Les plus bas taux de classification s'expliquent par des transitions entre états rares ou particulières, qui n'apparaissent pas dans l'ensemble d'apprentissage, comme, par exemple, des fondus enchaînés entre deux échanges. Les confusions les plus fréquentes ont lieu entre les "premiers services ratés" et les "échanges". L'entrelacement temporel des plans ne suffit pas toujours à les differencier. Les temps morts sont en revanche bien détectés.

Les résultats obtenus pour les séquences dont la famille a été totalement exclue de l'ensemble d'apprentissage, sont similaires aux autres. Cela prouve les capacités de généralisation du modèle.

Nous utilisons ensuite les résultats de la classification par la méthode 2 précédemment présentée. Les erreurs de classification se répercutent sur la segmentation et dégradent légèrement les performances de la segmentation.

	man.	auto.	sim.
US_Open_05	73	70	69
RGO2_set3	62	55	49
RG01_set1	73	55	65
DavisCup_set1	61	58	60
DavisCup_set2	55	54	57
DavisCup_set3	67	66	62
OpenParis_set1	63	56	60
moyenne	64	60	60

TAB. 2.6: Précision de la segmentation : "man." désigne les données manuellement annotées, "auto." celles issues d'une étape de classification automatique préalable, et "sim." la similarité visuelle.

2.6.3.2 Classification et segmentation simultanées

Dans cette approche, la classification en terme de "vues du terrain"/"autres vues" est réalisée simultanément à la segmentation en unités logiques. La classification est donc probabiliste. De plus, elle est guidée par le contexte. Le tableau 2.6 présente la précision de la segmentation sur l'ensemble des données de tests pour les trois approches proposées. Les performances sont identiques, que les données aient été préalablement classifiées ou non. Cependant dans le premier cas, il y a toujours le problème du seuil à fixer pour la

classification. D'autant plus que la segmentation en unités logiques est très sensible aux résultats de la classification. Une fois la décision sur la classe, il n'est plus possible de la corriger dans le processus de segmentation. L'état binaire associée à la classe est une contrainte très forte qui pénalise toute information supplémentaire. Pour toutes ces raisons, nous choisissons de représenter les observations par leur similarité visuelle.

Vidéos	Précision segmentation	Premiers services		Echanges		Rediffusions		Temps morts	
		P	R	P	R	P	R	P	R
US_Open_05	69	40	33	79	62	54	68	88	89
RGO2_set3	49	60	50	55	20	38	71	90	60
RG01_set1	65	53	57	60	44	60	65	84	86
DavisCup_set1	60	67	43	46	53	61	84	83	60
DavisCup_set2	57	63	46	54	64	58	78	78	70
DavisCup_set3	62	91	45	50	46	60	93	90	57
OpenParis_set1	60	61	50	44	41	75	88	75	66
moyenne	60	62	46	55	41	58	78	84	70

TAB. 2.7: Résultats de la segmentation en unités logiques avec une classification simultanée des plans en vues du terrain. P désigne la précision et R le taux de rappel.

Le tableau 2.7 détaille les taux de précision et de rappels obtenus pour chaque unité logique, pour toutes les séquences de test. Contrairement à ce que pourrait laisser penser les faibles taux de précision et de rappel pour les "premiers services" et les "échanges", la similarité visuelle permet de distinguer les vues globales des autres. En moyenne sur l'ensemble des séquences, les taux de précision et de rappel de la classification a posteriori en vues de terrain sont respectivement de 84% et 87%, et de 96% et 90% pour les autres vues. Ce sont les unités logiques relatives qui sont souvent confondues. La similarité visuelle repère bien une vue du terrain, mais l'analyse de l'entrelacement temporel des plans échoue. Les forts taux de précision des temps morts confirment la capacité du système à discriminer les vues du terrain des autres.

La détection des rediffusions se base essentiellement sur la détection des fondus enchaînés. En l'absence de fausses détections dans la segmentation, la contrainte *a priori* sur la présence des fondus enchaînés est très forte, et on observe 100% de bonnes classifications des rediffusions. Les données d'apprentissage sont cependant bruitées, et les fausses détections dans l'apprentissage polluent les probabilités de transitions. Des transitions non souhaitées *a priori* ont des probabilités non nulles. Lors du décodage, tous les fondus enchaînés détectés ont tendance à être classés comme rediffusion, ce qui explique le faible taux de précision associé à cette classe. Nous avons testé la segmentation sur une séquence pour laquelle nous disposions d'une segmentation en plans corrigée. La performance de la segmentation est alors de 76%, contre 65% dans le cas de la segmentation automatique. Il s'agit d'un gain très important. Notre approche étant basée plan, elle est très sensible à la qualité de la segmentation temporelle. Il est néanmoins encourageant de supposer que les performances du système augmentent d'environ 10% dès lors que la détection des transitions progressives devient fiable.

2.7 Conclusion

Nous avons défini des unités logiques d'une vidéo de sport dans le cas du tennis et du baseball. Ces unités logiques portent une information sémantique sur le statut du jeu. L'information *a priori* résultant de la syntaxe d'un match de tennis est prise en compte dans la construction de modèles syntaxiques. Nous avons notamment défini quatre éléments structuraux de base pour le tennis : deux d'entre eux sont relatifs à des phases de jeu (les "premiers services ratés" et les "échanges") et les deux autres sont relatifs à des phases de non-jeu (les "temps morts" et les "rediffusions").

L'unité temporelle utilisée est le plan vidéo. La segmentation en unités logiques réalise une macro-segmentation dense de la vidéo de plus haut-niveau qu'une segmentation en phase de jeu/phase de non-jeu.

À l'intérieur de chaque unité logique, un modèle de Markov caché représente l'entrelacement temporel des plans. Un état d'un HMM représente un plan de la vidéo. Chaque plan est caractérisé par son type de point de vue et sa durée. L'architecture des HMMs est définie *a priori* de façon à intégrer les règles de production. La classification et la segmentation en unités logiques sont réalisées simultanément par un algorithme de programmation dynamique (algorithme de Viterbi). Les attributs bas-niveau sont la similarité visuelle et la longueur des plans. Ces attributs sont indépendants de la vidéo étudiée.

La segmentation en unités logiques a été testée sur un ensemble de séquences variées. Les résultats montrent que les attributs visuels utilisés ne suffisent pas à déterminer toutes les scènes. Dans le chapitre suivant, nous étudions l'apport d'attributs audio supplémentaires.

Le travail présenté dans ce chapitre a fait l'objet d'une publication [89].

Chapitre 3

Intégration d'indices audio

Comme cela a été vu au chapitre 1, la plupart des solutions proposées pour l'indexation utilisent une approche unimodale. Nous nous intéressons dans ce chapitre à l'intégration dans le système précédant d'informations extraites du signal audio. Nous évoquerons tout d'abord les différentes méthodes d'intégration audiovisuelle, avant de présenter la méthode que nous avons mise en œuvre et la représentation des informations audio que nous avons choisie. Nous évaluerons enfin les performances de la segmentation audiovisuelle en unités logiques.

3.1 Introduction

Un document vidéo est composé de trois sources d'informations ou modalités :
- *modalité visuelle* : contient tout ce qui peut être vu dans le document ;
- *modalité audio* : contient la parole, la musique et les bruits d'ambiance qui peuvent être entendus dans le document ;
- *modalité textuelle* : contient les ressources textuelles qui décrivent le contenu du document. Notons que cette dernière modalité, connue sous le nom de *caption stream*, n'est pas disponible en Europe.

Nous avons passé en revue l'utilisation d'indices visuels pour les systèmes spécifiques au chapitre 1. La plupart des approches n'utilisent qu'une seule modalité. Les efforts actuels se dirigent cependant vers l'intégration des informations fournies par les différents médias.

Nous présentons tout d'abord les informations extraites du signal audio qui sont utilisées par les systèmes spécifiques d'analyse des événements sportifs, puis la façon dont la problématique de la multimodalité a été abordée dans ces systèmes.

3.1.1 Utilisation de la bande sonore

Bien que la majorité des systèmes spécifiques s'appuie sur l'analyse visuelle, quelques travaux exploitent uniquement l'information audio [90, 91, 92, 5]. L'utilisation de l'information audio est motivée par deux raisons. Primo, l'analyse du signal audio est plus rapide que celle du signal vidéo. Elle évite de recourir à des caractéristiques images coûteuses à extraire et à traiter. Secundo, le signal audio fournit plus directement des informations de haut-niveau sur le document.

Le signal audio associé aux retransmissions sportives télévisées est fortement bruité, ce qui rend son analyse difficile. Plusieurs sources sonores sont mélangées : le discours du commentateur, le bruit du public, le bruit du jeu : arbitre, impact de la balle, etc. Dans la plupart des cas, l'objectif de l'analyse audio est de séparer et d'identifier différentes classes de sons présentes dans le signal. On distingue trois catégories de classes sonores :

- génériques (communes à une majorité de documents audiovisuels) : il s'agit des classes parole, musique et silence ;
- spécifiques au domaine (communes à tous les sports) : acclamation de la foule ;
- spécifiques au sport : impact de la balle et de la raquette ou de la batte, impact du ballon sur le sol pendant les dribbles, coup de sifflet.

L'extraction de caractéristiques d'un signal pour la description de son contenu est généralement guidée par l'analyse acoustique. Les attributs du signal audio se décomposent en trois familles :

- les attributs issus du domaine temporel : *short-time energy*, statistiques sur l'énergie (moyenne, écart-type,...), *silence ratio*, passages par zéros (ZCR pour *Zero Crossing Rate*) et *pause rate* ;
- les attributs issus du domaine fréquentiel : fréquence fondamentale (*Pitch*), énergie des sous-bandes, coefficients cepstraux (MFCC pour *Mel Frequency Cepstral Coefficient*), coefficient de prédiction linéaire (LPC), statistiques spectrales : centroïde, largeur de bande...
- les attributs psycho-acoustiques : modulation d'énergie à 4Hz, fréquence de coupure spectrale.

Le lecteur pourra trouver les définitions détaillées, ainsi que les méthodes d'exploitation de ces attributs dans [93] pour le traitement de la parole et [94, 95, 96] pour l'indexation et l'analyse vidéo.

Une caractérisation bas-niveau du signal sonore particulièrement utilisée pour l'analyse des vidéos de sport consiste à détecter les acclamations de la foule et l'excitation dans la voix du commentateur [6, 97]. Ces deux événements sonores sont facilement caractérisables à partir des attributs spectraux et temporels du signal. Ces approches sont généralement basées sur l'analyse de l'énergie du signal. Elles ne permettent pas d'autre classification que "classe de haute énergie"/"classe de basse énergie" et sont souvent basées sur la définition d'un seuil variant d'une vidéo à l'autre.

Une caractérisation de plus haut niveau consiste à segmenter le signal en classes. La segmentation en classes à partir des attributs bas-niveau fait appel à des méthodes supervisées parmi lesquelles on peut citer les réseaux de neurones [98, 91], les machines à support de vecteurs [5, 90, 8], les modèles de Markov cachés [99], les mélanges de modèles de gaussiennes [100, 101] et les réseaux bayésiens [102].

L'analyse de la bande sonore amène une classification en action/non-action de la vidéo. La définition de l'action varie en fonction du sport, si on considère des événements spécifiques tels que l'impact de la balle ou les coups de sifflets.

Une hypothèse générale utilisée, valable quel que soit le type de sport, est que les événements intéressants de la vidéo sont caractérisés par une agitation du présentateur et de la foule, sous forme d'acclamations ou d'applaudissements. Par exemple, la détection des acclamations est utilisée pour détecter des événements au basketball [91], les applaudissements et l'excitation du commentateur pour le tennis et le football [92]. Sur le même

principe, Xiong *et al.* proposent une classification générique de la vidéo en action/non-action. La bande sonore est d'abord classifiée en applaudissements, acclamations, musique, parole, et mélange parole-musique. Les deux premières classes sont utilisées pour détecter les événements tandis que les trois dernières servent à filtrer les segments inintéressants. Pour prouver l'indépendance de cette technique vis à vis du sport, l'approche est validée pour le football, le golf et le baseball. La combinaison avec des indices plus spécifiques rend la détection d'événements plus robuste. Ainsi pour le baseball, la détection, d'une part, de l'impact de la balle et de la batte et, d'autre part, de l'excitation dans la voix du commentateur, sont fusionnées [90].

Ces techniques détectent les événements indépendamment de la mise en œuvre de modèles spécifiques. La détection des segments intéressants permet des applications de type résumé ou navigation rapide. Elles ne permettent pas en revanche la reconnaissance de ces événements. La détection d'événements par détection des acclamations et de l'excitation du commentateur est l'équivalent audio de la détection des rediffusions dans la vidéo.

L'identification des événements nécessite des caractéristiques plus spécifiques et l'utilisation de modèles temporels. De la même façon que le type et la succession des prises de vue dans la vidéo fournit une information sur le statut du jeu, la détection de différents types d'événements audio et l'analyse de leur succession permet d'identifier des événements particuliers. Xu *et al.* [8] identifient cinq événements (coup franc/penalty, poussée irrégulière, but, tir au but, début/fin jeu) dans une vidéo de football en combinant l'interprétation de certains types d'événements sonores (voir table 3.1) avec des règles de décision heuristiques du type :

- si l'intervalle de temps entre un double coup de sifflet et un long coup de sifflet est inférieur à un certain seuil alors il s'agit d'un coup-franc ou d'un penalty
- si il y a un double coup de sifflet alors il s'agit d'une poussée irrégulière

Evénement audio	Interprétation sémantique
coup de sifflet long	début d'un coup-franc, penalty ou d'un corner
double coup de sifflet	poussée irrégulière
coup de sifflet multiple	rappel de l'arbitre
excitation voix commentateur	but ou tir au but
excitation du public	moments excitants
parole commentateur	normal

TAB. 3.1: Interprétation sémantique des événements audio pour le football d'après [8].

3.1.2 Problématique de la multimodalité

Il est tout à fait raisonnable de penser que l'intégration de plusieurs sources d'informations augmente les performances de l'analyse du contenu d'une vidéo, comme l'indiquent déjà quelques travaux [103, 17].

L'intégration audiovisuelle en particulier est étudiée dans différents domaines de l'analyse du contenu : reconnaissance de la parole, où les données visuelles sont les contours de la bouche [104, 105] ou des paramètres d'animation faciale [106], et pour différents

domaines de l'indexation vidéo, comme la segmentation temporelle [107, 108, 109], l'identification d'unités logiques dans les journaux télévisés [77, 110, 111, 112], la détection des dialogues [113, 114, 115], la création de résumés vidéo [116, 117], la classification en genre [118, 119] et la détection de concepts sémantiques comme le "lancement d'une fusée" [120].

Cependant l'intégration des caractéristiques issues de différentes modalités n'est pas une tâche triviale et nous pouvons souligner au moins deux problèmes parmi d'autres :

un problème de décision commun à tous les systèmes de fusion d'informations : quelle doit être la décision finale lorque les différents médias fournissent des informations contradictoires ?

un problème de synchronisation propre à l'intégration multi-modale : la fréquence d'échantillonnage qui permet de calculer et d'analyser les attributs bas-niveau n'est pas la même selon les médias :

- l'unité élémentaire du signal vidéo est l'image ; à une fréquence d'échantillonnage de 25Hz, il est possible d'extraire un attribut image toutes les **40 ms** ;
- l'unité élémentaire du signal audio est la trame ; à une fréquence d'échantillonnage de 100Hz, il est possible d'extraire un attribut audio toutes les **10 ms**. Pour une analyse statistique, l'unité utilisée par de nombreux auteurs est le clip qui est un ensemble de trames consécutives d'une durée de 1 à 3 secondes. Dans cet intervalle de temps, le signal audio est considéré quasi-stationnaire.

Bien entendu, les frontières d'un découpage temporel du signal visuel en plans, et les frontières d'un découpage du signal audio en segments homogènes ne coïncident pas. Les approches qui s'intéressent à la segmentation en scènes de la vidéo doivent étudier le voisinage de chaque frontière pour estimer s'il y a un changement conjoint.

Pour résoudre ces deux problèmes, les informations auditives et visuelles peuvent être combinées de différentes façons. La première d'entre elles repose sur l'utilisation successive des analyses audio et vidéo. Une autre consiste à combiner les attributs audio et vidéo au sein d'un unique vecteur de caractéristiques audiovisuelles avant la classification. On parle alors d'intégration précoce (*early integration*). Enfin, la dernière consiste à faire deux classifications indépendantes selon chaque modalité, puis à fusionner leur résultats. Il s'agit de l'intégration tardive (*late integration*).

La mise en application de ces trois approches dans le cadre de l'analyse des vidéos de sport est développée dans les paragraphes suivants.

3.1.2.1 Analyse successive

Le principe de l'analyse successive est le suivant : le signal audio ou textuel est d'abord utilisé pour détecter les segments intéressants. L'analyse de l'image (suivi, détection des lignes, des contours, segmentation spatiale) est ensuite utilisée dans les régions précédemment sélectionnées afin d'identifier un événement particulier [98, 121], ou plus simplement pour localiser les frontières vidéo du segment [122]. Dans le premier cas, l'audio ou le texte sont utilisés pour restreindre la fenêtre temporelle pour l'analyse vidéo. Il s'agit, en fait, d'une méthode de prédiction/vérification. Pour être intéressante, cette approche suppose que la détection des segments intéressants peut se faire plus rapidement avec les attributs audio. Ces derniers sont donc généralement de bas-niveau (acclamation de la foule). Le flux

textuel est utilisé pour détecter des mots-clés pré-définis identifiant l'événement probable. La vérification et la localisation sont effectuées sur les flux audio et/ou visuel [121, 123].

Le schéma inverse est également proposé dans [28, 124] : dans un premier temps, les caractéristiques visuelles sont utilisées pour détecter les phases de jeu, ou des événements intéressants. Dans un deuxième temps, la mesure de l'excitation du commentateur ou du public sélectionne les plans les plus intéressants. Dans ce sens, il ne s'agit plus d'un schéma de prédiction suivi d'une vérification. Le signal sonore est utilisé en complément pour ordonner les segments visuels candidats par importance.

Lorsque ce sont les caractéristiques audio qui sont utilisées pour identifier les événements, elles doivent être de plus-haut niveau, donc plus difficiles à extraire. Afin de limiter la dégradation de l'extraction des attributs liée à la présence de bruit, les segments audio à analyser peuvent être présélectionnés d'après l'analyse du signal vidéo. Xu *et al.* [5] s'intéressent à l'identification des événements dans une vidéo de tennis par une analyse audio des plans de jeu. Ces derniers sont préalablement identifiés par une classification visuelle (figure. 3.1). Ainsi, l'analyse audio n'est réalisée que pour les plans supposés contenir des bruits de balle.

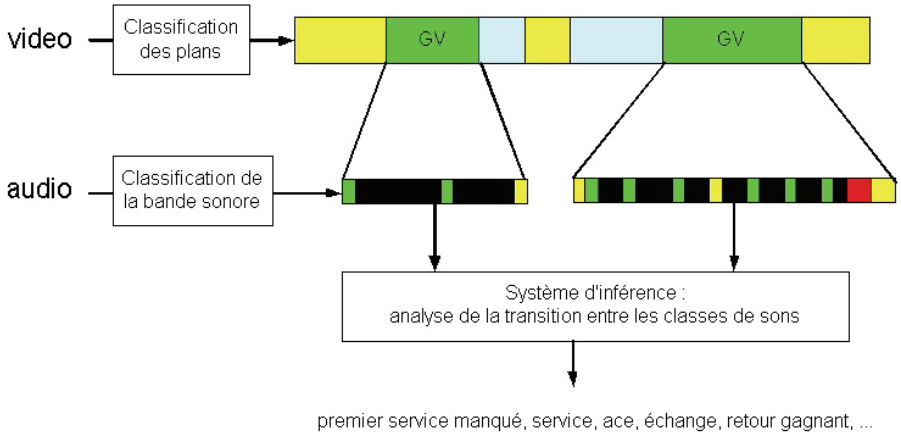


FIG. 3.1: Schéma de l'analyse audio-visuelle d'une vidéo de tennis [5].

La bande sonore est segmentée en quatre classes : bruit de balle, silence, applaudissements et parole. Les événements sont caractérisés par des motifs de transitions particuliers entre les classes sonores considérées, représentés à la figure 3.2. Ils sont identifiés en calculant l'intervalle de temps entre deux bruits de balle, et en vérifiant la présence des applaudissements à la fin d'un plan.

Puisque nous nous intéressons particulièrement au tennis, citons une dernière approche de combinaison successive des informations audio et vidéo, bien que l'objectif soit un peu en marge des travaux précédemment présentés. En analysant la position des joueurs et de la balle au moment de l'impact entre la raquette du joueur et la balle, Miyamori [39] identifie les actions des joueurs, comme les coups droit, les revers et les smash. Le moment de l'impact ne peut être estimé correctement uniquement par des méthodes visuelles de suivi de la balle et des joueurs, notamment à cause des inévitables occultations qui ont lieu au moment de l'impact, la balle étant par définition très proche du joueur. Le moment de

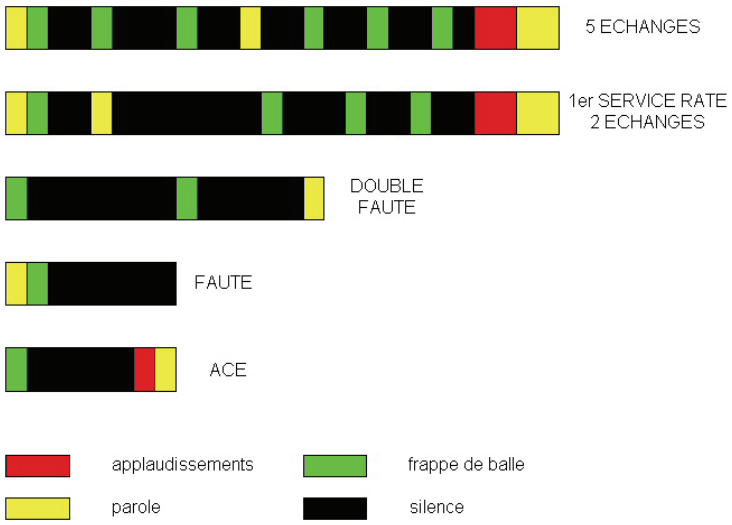


FIG. 3.2: Cinq motifs de transitions entre les classes sonores caractéristiques du tennis, à l'intérieur d'un plan du terrain, et leur signification sémantique [5].

l'impact t_i est alors déterminé en analysant la bande sonore par une méthode de template-matching dans le domaine de Fourier. Il s'agit d'une méthode très coûteuse en temps de calcul, tant du point de vue de l'analyse du signal sonore que de celui du signal vidéo, compte-tenu de l'objectif final de reconnaissance d'une action humaine.

3.1.2.2 Intégration précoce

L'intégration précoce consiste à intégrer les attributs audio et vidéo au sein d'un même vecteur avant la classification. Cette forme d'intégration nécessite de synchroniser les données. La synchronisation des données peut se faire à différents niveaux et s'aligner sur l'une ou l'autre des modalités :

- niveau frame audio (0.01s) [102]. Les attributs images sont alors sur-échantillonnés par interpolation, ou simplement répétés ;
- niveau clip audio (1s) [100, 125]. Les attributs vidéo sont alors moyennés sur cette durée ;
- niveau image [126] ;
- niveau plan vidéo [127] ;

La synchronisation au niveau du plan est utilisée pour la classification des plans d'une vidéo de baseball [101] et pour la détection d'événements au baseball [127]. Pour chaque plan, les informations sonores, visuelles et textuelles sont combinées dans un vecteur de caractéristiques. Dans le premier cas (classification des plans), la variation temporelle des attributs durant un plan est prise en compte en subdivisant le plan en trois parties de longueurs égales. Dans le deuxième cas (détection d'événements), ce sont les informations contextuelles liées aux plans voisins qui sont prises en compte. Un modèle est construit pour chaque classe de plans ou d'événements, et la classification est réalisée pour chaque plan à partir du vecteur de caractéristiques, par la méthode du maximum d'entropie.

Au lieu d'un vecteur de caractéristiques, les informations audiovisuelles peuvent être combinées au sein d'un histogramme [100]. Les données sont représentées au niveau symbolique. Chaque clip audio est labelisé selon l'une des classes sonores prédéterminées. L'intensité du mouvement est quantifiée et synchronisée sur la durée d'un clip audio. Finalement, un histogramme 2-D joint des labels mouvement et audio est calculé sur une fenêtre temporelle. Les événements sont détectés par une mesure d'entropie entre deux fenêtres.

3.1.2.3 Intégration tardive

Dans cette approche, chaque modalité est classifiée indépendamment. L'intégration est réalisée au niveau de la prise de décision. Celle-ci peut être basée sur un ensemble de règles heuristiques. Par exemple, l'audio et la vidéo sont segmentés et classifiés par deux modèles de Markov distincts. Les dialogues sont identifiés comme des scènes pour lesquelles le signal audio est majoritairement de la parole, et pour lesquelles l'information visuelle apparaît de façon alternée. La détection de ces scènes particulières est réalisée par fusion des décisions [114].

Les règles heuristiques sont également utilisées pour détecter des paniers dans une vidéo de basketball [97]. Les détections des acclamations dans le signal audio, du texte incrusté et du changement de mouvement de la caméra dans la vidéo sont d'abord réalisées indépendamment. L'identification des paniers se fait à partir de modèles temporels des attributs détectés, basés sur des règles heuristiques.

Pour détecter les plans de jeu d'une vidéo de tennis, Dahyot *et al.* [30] définit indépendamment la vraisemblance visuelle d'un plan d'être une vue du terrain, et la vraisemblance d'un segment audio de représenter l'impact de la balle. La décision finale est prise en seuillant le produit des deux vraisemblances.

Hanjalic [126] détecte les buts en essayant de caractériser l'excitation contenue dans la vidéo. Pour cela, l'activité, une mesure de densité des coupures et l'énergie de la bande sonore sont calculées pour chaque image. Puis les courbes temporelles de ces trois attributs sont tracées, afin de détecter des pics. Comme ceux-ci ne sont pas nécessairement simultanés, les maxima locaux de chacune des courbes dans une fenêtre temporelle sont fusionnés.

Pour résoudre ce problème de non-simultanéité des événements dans les différentes modalités, une approche alternative a été présentée dans [6]. Les relations d'intervalles de temps entre les différents attributs détectés, comme "précède" et "pendant" sont explicitement modélisées (figure 3.3). La détection des événements est traitée comme un problème de reconnaissance des formes. Pour un plan vidéo i donné, la forme étudiée est le vecteur des n attributs extraits dans les différents modalités et de leur relation avec le plan i. Un modèle du type : "un but est un plan précédé d'une translation de caméra et de texte incrusté, et terminé par l'excitation du commentateur" est crée pour chaque événement. La classification du plan est réalisée par le maximum d'entropie.

3.1.3 Multimodalité et modèles de Markov cachés

Nous venons de voir un panorama de différentes méthodes existant pour intégrer des informations issues de médias différents. Trois approches ont été proposées pour l'analyse multimodale : l'analyse successive des différents médias, les méthodes d'intégration pré-

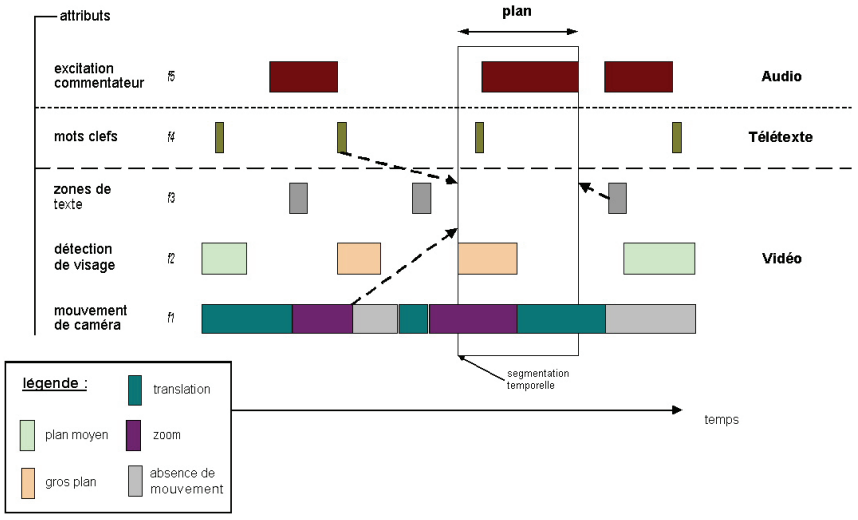


FIG. 3.3: Représentation de la relation contextuelle (flèches) entre un événement et les attributs d'un document vidéo d'après [6].

coce et les méthodes d'intégration tardive. Nous nous intéressons ici aux deux dernières méthodes et à leur implication dans une approche utilisant les modèles de Markov cachés.

L'intégration précoce est une méthode simple mais coûteuse en temps de calcul. La complexité de calcul augmente avec la taille de l'espace des attributs, de même que le nombre de données nécessaires pour l'apprentissage. Si V est la dimension du vecteur de caractéristiques visuelles, et A la dimension du vecteur de caractéristiques audio, du point de vue des modèles de Markov cachés, la concaténation des vecteurs audio et vidéo se traduit par un HMM mono-flux, pour lequel les lois de probabilités sont estimées dans un espace de taille $A + V$. Cette approche nécessite également de synchroniser les caractéristiques.

Une alternative à ces inconvénients est l'intégration tardive, au niveau de la décision. Cette dernière ne tient cependant pas compte des dépendances entre les attributs des différentes modalités. L'intégration au niveau de la décision se traduit quant à elle de deux façons, du point de vue des HMMs. Soit les états audio et vidéo sont synchronisés, et on parle alors de HMM multi-flux. Les deux flux sont utilisés séparément pour modéliser les informations audio et vidéo. Les HMMs multi-flux autorisent l'audio et la vidéo à apporter des contributions différentes à la probabilité d'observation. Soit les états sont asynchrones, et on parle alors de HMMs produits, qui sont des extensions des HMMs multi-flux dont la topologie autorise l'asynchronisme entre les états (figure 3.4). Il s'agit de deux HMMs indépendants, l'un pour l'audio, l'autre pour la vidéo. Les vraisemblances de chaque séquence d'observations sont combinées en fonction de la confiance de chaque modalité.

Nous avons choisi d'utiliser des HMMs multi-flux synchrones. D'une part, cette méthode conserve la topologie des HMMs, et donc la modélisation des unités logiques réalisée pour le flux visuel. D'autre part, cette approche intégre les informations sonores de façon simple et directe au schéma existant. Dans le paragraphe suivant, nous allons détailler l'approche que nous proposons et les attributs audio que nous utilisons.

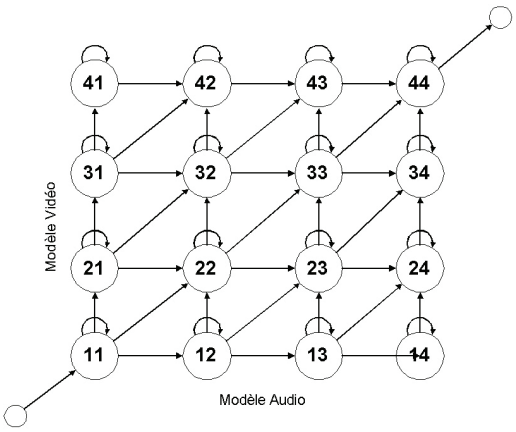


FIG. 3.4: HMM produit.

3.2 Intégration dans le modèle de Markov

3.2.1 Méthode proposée

Nous proposons ici une méthode intermédiaire entre l'intégration au niveau des attributs et l'intégration au niveau des décisions. Techniquement, cette approche consiste à réduire l'espace des attributs en prenant un certain nombre de décision indépendamment sur chacune des modalités.

Sur un autre niveau, l'enjeu est de savoir si les observations audio apportent de l'information utile pour le système. On décide de représenter celles-ci par des attributs de haut niveau. Si les performances du système ne sont pas améliorées, la cause ne sera pas clairement attribuable à la forme trop simple des représentations du signal audio, ou au fait que celui-ci ne véhicule aucune information supplémentaire. En revanche, si les performances s'améliorent, cela démontre que le signal audio amène un supplément d'information.

Aucune décision sur la classe visuelle n'est prise avant l'intégration. En revanche, un certain nombre de décisions dures sont prises sur le signal audio bas niveau avant l'intégration. Le signal audio est finalement représenté par un vecteur binaire décrivant la présence de certaines classes dans le plan vidéo. Nous avons en effet choisi le plan vidéo comme unité temporelle d'intégration.

L'architecture du système est décrite dans la figure 3.5. Les flux audio et vidéo sont traités séparément pour extraire les observations.

Dans la partie suivante, nous décrivons la nature de l'observation audio que nous utilisons et son intégration dans le système existant. Puis nous présentons les résultats expérimentaux très encourageant de la segmentation audio-visuelle en unités logiques. Cela valide à la fois la représentation du signal sonore choisie et l'apport par le media audio d'informations complémentaires dans notre application.

3.2.2 Description des attributs audio utilisés

Pour chaque plan vidéo, un vecteur binaire a_t décrit quelles classes, parmi {"parole", "applaudissements", "tennis" (bruit de balle), "bruit", "musique"}, sont présentes dans le

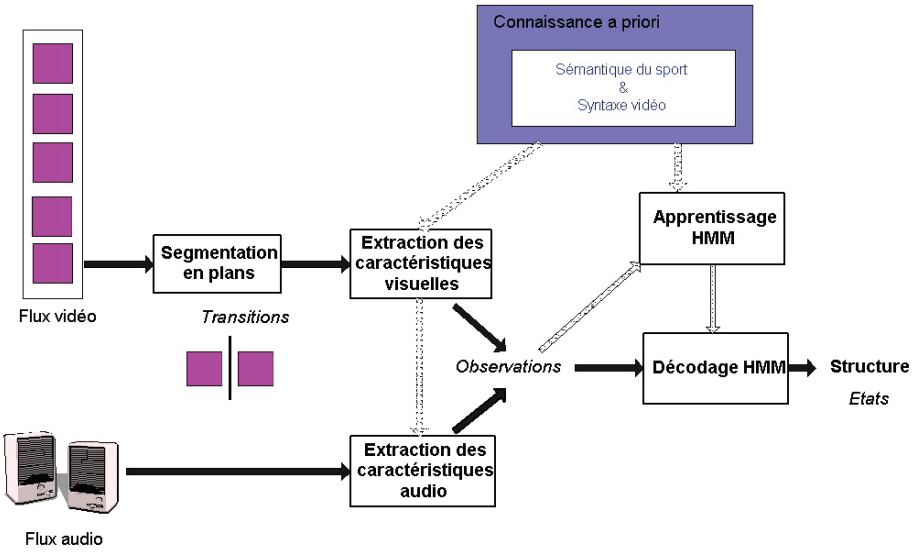
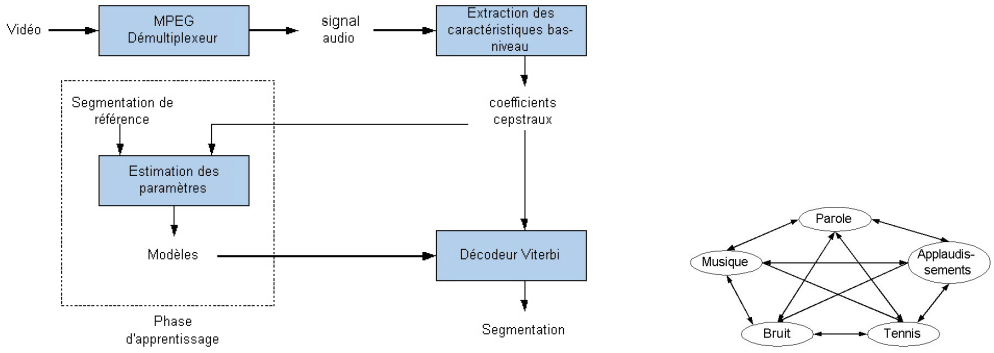


FIG. 3.5: Architecture du système de structuration audiovisuel.

plan vidéo.

L'extraction des attributs audio a été réalisée par l'équipe METISS de l'IRISA. Le lecteur pourra trouver des informations plus complètes sur la méthode mise en œuvre dans [128]. Nous en exposons ici les grandes lignes.

Le vecteur binaire est extrait à partir d'une segmentation automatique de la bande sonore. Le signal audio est démultiplexé de la vidéo et converti en descripteurs représentatifs du contenu. La segmentation est réalisée par un algorithme de Viterbi associé à un HMM ergodique, dans lequel chacun des états représente une classe audio C_i (figure 3.6). Les densités de probabilité $p(x_i|C_i)$ sont des mélanges de Gaussiennes dont les paramètres sont estimés lors d'une phase d'apprentissage. L'introduction d'une classe "bruit" diminue les erreurs de classification inévitables lorsqu'un ensemble fini de classes est utilisé et que le signal ne correspond pas à l'une d'entre elles.



(a) Architecture du système (b)Modèle de Markov correspondant

FIG. 3.6: Segmentation et classification du signal sonore.

	musique	parole	tennis	appl.	bruit	suppr.
musique	16.49	14.35	17.13	13.33	1.59	37.11
parole	0.11	87.15	2.21	0.67	1.63	8.23
tennis	0	5.67	77.16	2.54	4.90	9.73
appl.	0.02	0.30	2.33	88.47	0.58	8.29
bruit	0.93	12.08	30.60	15.52	40.87	0.00
ins.	3.3	2.96	65.93	14.11	0.00	

TAB. 3.2: Matrice de confusion de la classification audio.

Les descripteurs audio considérés sont les coefficients cepstraux et le logarithme de l'énergie, calculés sur des fenêtres consécutives de 20ms se recouvrant à 50%. Les 16 coefficients cepstraux et l'énergie sont complétés par les dérivées de premier et de second ordre.

Chacune des 5 classes audio est modélisée par un mélange de 64 gaussiennes. Une phase d'apprentissage est réalisée sur des données manuellement annotées, afin d'estimer les paramètres des modèles de mélanges. L'un des problèmes rencontrés par la segmentation audio de la bande sonore est que plusieurs événements peuvent être présents simultanément. Par exemple, il est courant d'avoir de la parole mélangée avec des bruits de balle ou des applaudissements. De tels événements composés ne peuvent être détectés en utilisant une segmentation reposant sur l'algorithme de Viterbi, à moins qu'un modèle soit donné pour chaque combinaison possible de classes. Etant donné que l'estimation des paramètres pour de tels modèles exigerait une très importante quantité (indisponible) de données, il est supposé qu'au maximum 2 événements peuvent être présents simultanément. Pour chaque paire d'événements (C_i, C_j), des modèles joints sont déduits des modèles des classes simples de la façon suivante :

$$p(x_t|C_i, C_j) = \frac{1}{2} p(x_t|C_1) p(x_t|C_2) \tag{3.1}$$

où x_t correspond au vecteur d'attributs de la trame t. L'hypothèse sous-jacente de ce modèle est que les classes C_i et C_j sont présentes dans le vecteur x_t avec une égale probabilité *a priori*. La densité $p(x_t|C_i, C_j)$ est un mélange de gaussiennes qui correspond à la concaténation des mélanges de gaussiennes de deux classes C_i et C_j, moyennant une renormalisation des poids.

La segmentation et la classification de la bande sonore sont finalement réalisées en utilisant un décodage de Viterbi, pour un HMM ergodique dans lequel chaque état représente soit une classe audio simple, soit une combinaison de 2 classes. Le taux de classification correcte obtenu est 77.8%, et le taux d'erreur de classification est 34.4%, dû aux insertions. La matrice de confusion correspondante est représentée dans la table 3.2. Les applaudissements, la parole et le tennis sont bien classifiés. Cependant le bruit est souvent classifié comme "tennis", probablement parce sur la classe "tennis" se compose d'un mélange de bruit de balle et de silence. La classe "tennis" est également souvent inserée comme le montre la dernière ligne de la matrice.

Finalement, les vecteurs audio a_t sont créés en regardant quels événements sont détectés par la segmentation audio, à l'intérieur des frontières fournies par la segmentation vidéo

(figure 3.7). La détection d'un événement se fait sur la base du rapport de vraisemblance entre la présence et l'absence de l'événement, le rapport de vraisemblance étant ensuite comparé à un seuil.

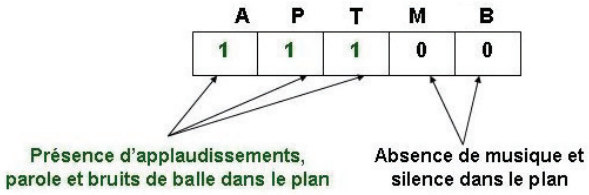


FIG. 3.7: Vecteur audio binaire décrivant les événements sonores apparaissant durant un plan vidéo.

3.2.3 Probabilité jointe

Pour chaque plan, les symboles d'observation issus de l'analyse de la vidéo sont ceux présentés au chapitre précédant : la durée du plan (cf. 2.4.1.1) et la similarité visuelle entre l'image-clé du plan et l'image-clé de référence K_{ref}, représentatives d'une vue globale (cf. 2.4.2). Nous y intégrons maintenant le vecteur audio qui caractérise la présence ou l'absence dans le plan d'événements sonores prédéterminés.

Formellement, pour un plan à l'instant t, l'observation o_t consiste donc en la similarité v_t, la durée du plan l_t et le vecteur de description audio a_t.

La probabilité de l'observation o_t d'être dans l'état j à l'instant t est alors donnée par :

$$b_j(o_t) = p(v_t|j)\ \ p(l_t|j)\ \ P[a_t|j] \tag{3.2}$$

où $P[a_t|j]$ est le produit sur l'ensemble des classes sonores k de la probabilité discrète $P[a_t[k]|j]$.

3.3 Résultats expérimentaux

Les expérimentations sont menées sur les 8 séquences pour lesquelles nous disposons de l'information audio. Trois d'entres elles sont réservées à l'apprentissage et les cinq autres composent l'ensemble de test. Pour chaque séquence, nous disposons de deux formes d'informations audio. La première est constituée des vecteurs audio issus d'une classification manuelle de la bande sonore. Il s'agit donc de données fiables destinées à évaluer l'apport de l'information audio sous la forme de vecteur binaire. Ces caractéristiques seront notées "man." dans les tables de résultats. La deuxième forme est constituée des vecteurs audio issus de la segmentation et de la classification automatique décrites plus haut. Ces données sont bruitées. Elles permettent de mesurer l'impact réel de l'information audio dans le système. Ces caractéristiques seront notées "deco." dans les tables de résultats.

3.3.1 Indices audio seuls

Nous considérons ici les informations audio seules, couplées à la durée des plans. La similarité visuelle n'est pas utilisée.

La synchronisation des données audio sur les frontières des plans vidéo est une méthode un peu brutale, dans le sens où les frontières de la segmentation audio en classes homogènes ne coïncident pas nécessairement avec les frontières des plans. Supposons que des applaudissements durent de la fin d'un échange jusqu'au début de l'échange suivant. Pour ce dernier plan, des applaudissements seront détectés alors qu'ils caractérisent les plans précédents. Ceci est notamment dû au fait que le vecteur audio binaire ne tient compte ni l'information temporelle des occurrences des événements audio, ni de leur importance dans le plan.

Pour pallier ce problème simplement, nous avons défini un taux de rejection T_{rej} tel que :

$$\text{si } \frac{\text{temps de présence de l'événement audio dans le plan}}{\text{durée du plan}} < T_{rej} \tag{3.3}$$

alors, l'événement audio n'est pas considéré dans le vecteur binaire.

Pour une séquence donnée, nous avons fait varier ce seuil de 0 à 60%. La figure 3.8 représente les taux de classification correcte de la segmentation, pour les vecteurs audio fiables (man.) et les vecteurs audio segmentés automatiquement (deco.) lorsque T_{rej} varie. Le taux de classification de la séquence en utilisant la similarité visuelle est donné en réference : 65%.

Les vecteurs audio fiables présentent des performances similaires aux attributs visuels pour $T_{rej} = 0\%$, et meilleures pour $T_{rej} = 10\%$. Dans la suite, nous utiliserons $T_{rej} = 10\%$.

En revanche, les résultats montrent que les erreurs de la segmentation automatique se répercutent sur le processus de structuration. Une autre source d'erreurs est l'inadéquation entre les données d'apprentissage (provenant de la vérité terrain) et les données utilisées pour le décodage (issues de la classification automatique). Comme le montre la figure 3.9, les résultats pour les vecteurs audio automatiquement segmentés sont meilleurs si l'apprentissage a été effectué sur des données issues elles aussi de la segmentation automatique.

Le tableau 3.3 compare les résultats de la classification en unités logiques lorsque les attributs visuels et audio sont utilisés séparément.

	visuel		deco.		man.	
	P	R	P	R	P	R
1^{er} service	53	57	37	55	57	67
Echange	60	44	46	34	80	76
Rediffusion	61	65	71	47	48	80
Temps mort	84	86	75	74	99	65

TAB. 3.3: Comparaison des résultats de la classification utilisant les attributs visuels seuls, les attributs audio segmentés automatiquement seuls et les attributs audio fiables seuls, pour la séquence RLDAMES set1.

Concernant les résultats de la classification utilisant uniquement des vecteurs audio fiables, l'augmentation des taux de rappel et de précision pour les échanges (resp. 80% et

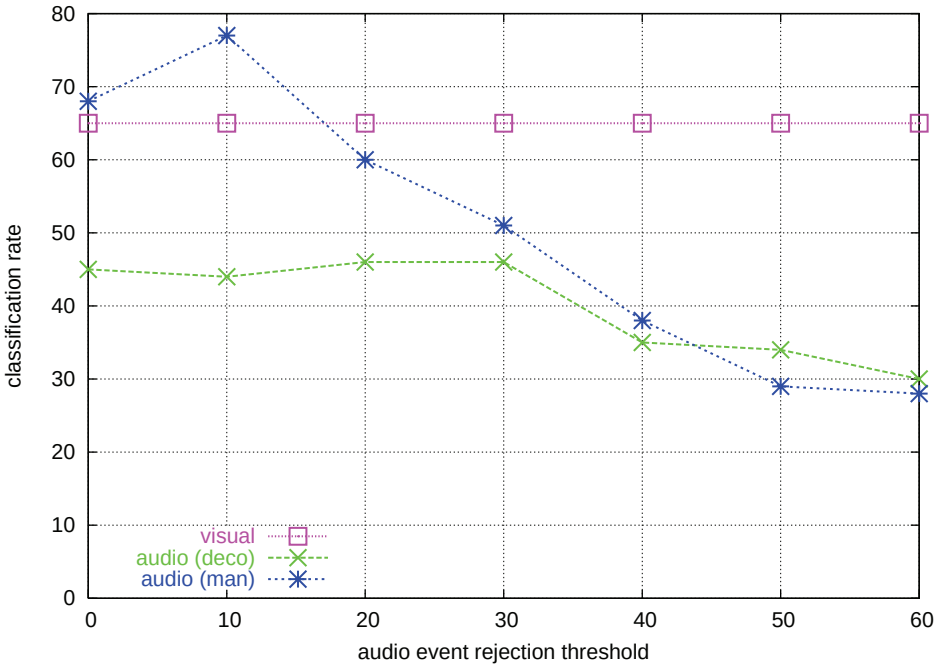


FIG. 3.8: Taux de classification correcte en utilisant les vecteurs audio fiables (man.) et les vecteurs audio segmentés automatiquement (deco.) lorsque T_{rej} varie de 0 à 60%.

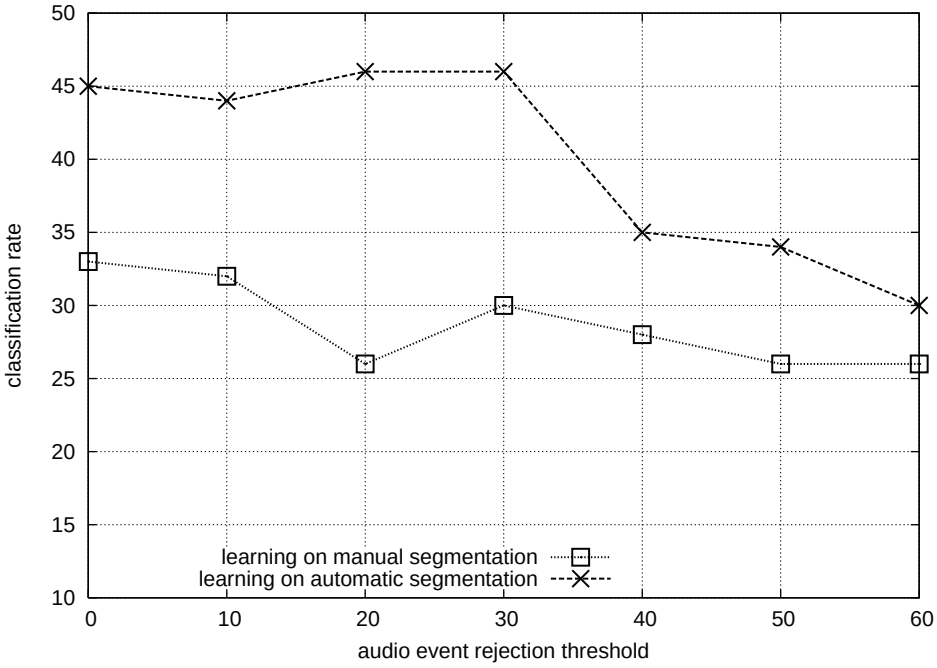


FIG. 3.9: Performances de la classification avec les vecteurs segmentés lorsque l'apprentissage a été réalisé sur des données fiables et sur des données décodées.

76%) montrent que les caractéristiques audio décrient efficacement les plans d'échanges. En effet, un échange se caractérise par la présence de bruits de balle et d'applaudissements, tandis que les premiers services ratés sont seulement caractérisés par la présence de bruits de balle. L'analyse de la bande sonore lève donc une ambiguïté entre les échanges et les premiers services ratés. Elle aide ainsi l'analyse de l'entrelacement temporel des plans. En revanche, les rediffusions, qui reposent essentiellement sur la détection des transitions progressives, ne sont pas caractérisées par un contenu audio représentatif, et par suite sont souvent mal identifiées (précision 48%). Les temps morts sont les seuls états pour lesquels l'événement "musique" (qui se produit pendant les publicités) est présent. La présence ou l'absence de musique dans un plan diminue donc le taux de mauvaises classifications. Cependant, tous les plans relatifs à des temps morts ne contiennent pas nécessairement de publicités donc de musique. L'absence de musique dans un tel plan augmente la probabilité de le manquer (rappel 65%).

Les résultats de la classification utilisant les vecteurs audio automatiquement segmentés sont en revanche plutôt mauvais. Le taux de rappel obtenu par la segmentation automatique de la bande sonore décrite en 3.2.2 est de 77.8%, et la précision est de 34.4%. Les événements "bruits de balle", "parole" et "applaudissements" sont bien classifiés, tandis que le bruit est souvent étiqueté comme "bruit de balle', probablement parce que la classe "bruit de balle" est caractérisée par un mélange de frappes de balles et de courts silences. La présence de bruits de balle ou de musique est respectivement synonyme d'échanges et de publicités. Dans l'ensemble des données de test, ces classes souffrent de nombreuses fausses détections. Les erreurs issues de la segmentation automatiques se répercutent au niveau de la structuration et dégradent les performances.

3.3.2 Indices audio et visuels

Nous intégrons cette fois les vecteurs audio extraits avec T_{rej} et les attributs visuels. L'intégration de ces deux indices augmente significativement les performances de la segmentation et de la classification. Les résultats de la segmentation audiovisuelle en unités logiques sont présentés dans le tableau 3.4 pour une séquence donnée. La détection des bruits de balle améliore le taux de classification des échanges, et la détection de la musique diminue les mauvaises classifications des temps morts. Par exemple, une vue globale n'est pas nécessairement représentative d'un échange (imaginons un échange fini toujours filmé par une vue globale). Elle est simplement le résultat de la volonté du producteur de montrer une vue globale du terrain à un instant donné. Dans un tel cas, les informations visuelles vont introduire une fausse détection d'un premier service raté.

	deco.		man.	
	P	R	P	R
1^{er} service	70	73	71	73
Echange	73	73	88	84
Rediffusion	70	77	63	87
Temps mort	90	84	89	78

TAB. 3.4: Résultats de la segmentation audiovisuelle en unités logiques sur la séquence RG01 set1.

Le tableau 3.5 présente les taux de classification audiovisuelle pour l'ensemble des séquences de tests. Les résultats sont légèrement meilleurs pour les classifications utilisant les vecteurs audio fiables. Cependant, l'apport de l'audio même bruité est important. L'augmentation des taux de classification par intégration audiovisuelle est une tendance manifeste sur l'ensemble des séquences.

	visuel	deco.	man.
DavisCup_set1	60	64	65
DavisCup_set2	57	66	69
DavisCup_set3	62	62	66
OpenParis_set1	60	74	78
RG01_set1	65	76	77
moyenne	61	68	71

TAB. 3.5: Comparaisons des performances de la segmentation lorsque les indices visuels sont utilisés seuls, ou couplés aux attributs audio.

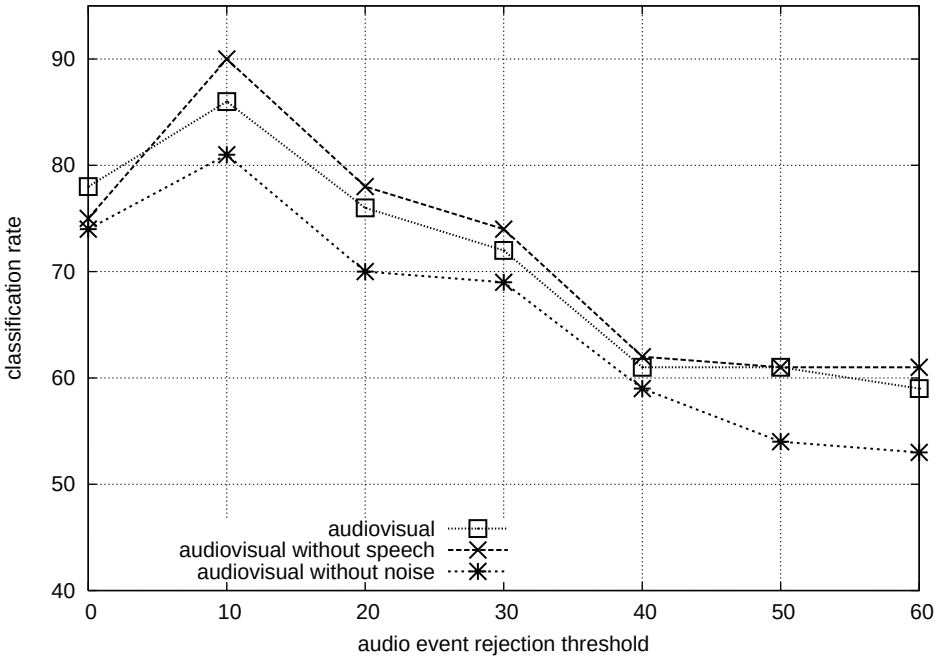


FIG. 3.10: Comparaison des taux de classification lorque les attributs visuels et audio sont utilisés conjointement et en supprimant la classe bruit et la classe parole.

Il découle de l'analyse précédente que les événements audio "bruits de balle", "musique" et "applaudissements" sont les plus discriminants. La figure 3.10 montre les taux de classification audiovisuelle lorque les événements "bruit" et "parole" ont été tour à tour supprimés des vecteurs audio fiables. Ne pas considérer la parole augmente légérement les performances. En effet, la parole est un événement présent dans quasiment tous les vecteurs

audio, avec une probabilité égale, du fait des commentaires permanents des commentateurs. Il ne devient pas de ce fait, un événement discriminant. En revanche, la suppression du "bruit" diminue les performances. A la différence de la classe "parole", le bruit n'a pas la même importance selon les classes. Par exemple, la probabilité de l'événement "bruit" dans un plan d'échange est supérieure à la probabilité de ce même événement dans les plans de temps mort.

3.4 Conclusion

Les HMMs fournissent donc également un cadre probabiliste efficace pour l'intégration de données multimodales. Les expérimentations montrent que l'intégration d'informations multimodales augmente les performances de la classification lorsque les caractéristiques utilisées sont de bonne qualité. En effet, les erreurs en provenance de l'extraction des caractéristiques se répercutent au niveau de la structuration. Il est donc nécessaire d'améliorer les performances du processus de segmentation de la bande sonore en classes. De plus, la description sonore en terme d'absence ou de présence d'une classe audio dans le plan est un modèle de représentation très simple qui peut être amélioré. Par exemple, il n'y a pas de mesure qualifiant l'importance de la présence de classe dans les plans. L'évolution temporelle du signal audio dans un plan peut aussi être prise en compte.

Le travail présenté dans ce chapitre a fait l'objet d'une publication [129].

Chapitre 4

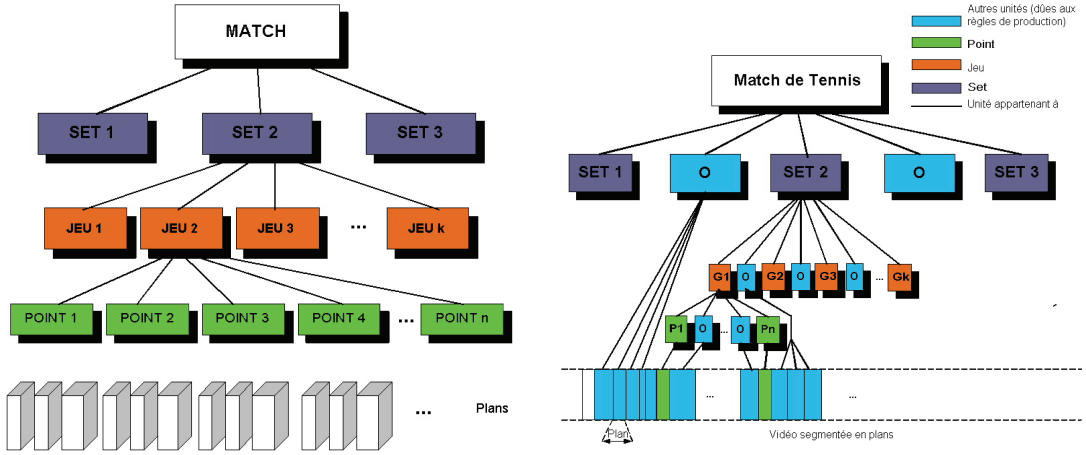
Modélisation par modèles de Markov cachés hiérarchiques

Dans ce chapitre, nous nous intéressons à la modélisation de la structure hiérarchique des événements sportifs. Les unités logiques précédemment définies sont intégrées dans un modèle de Markov caché hiérarchique décrivant la structure d'un match. Nous introduisons les modèles de Markov cachés hiérarchiques avant de les appliquer à la modélisation de la structure d'un match. Nous expliquons pourquoi cette modélisation nous amène à définir un nouveau symbole d'observation.

4.1 Introduction

Nous nous intéressons à des événements sportifs présentant une forte structure intrinsèque. Cette structure liée à la contrainte sur les scores est hiérarchique, comme le représentent les figures 4.1 pour le tennis et 4.2 pour le baseball. L'objectif de cette partie est d'identifier automatiquement chaque niveau de la hiérarchie dans les vidéos. Par exemple pour le tennis, il s'agit d'identifier l'ensemble des plans correspondant au premier set, puis le sous-ensemble correspondant au premier jeu, etc. A partir de la segmentation en plans de la vidéo et de l'analyse de l'entrelacement temporel des plans, nous avons déjà identifié des groupes de plans appartenant à des unités logiques. La succession de ces unités logiques est étroitement liée au déroulement du jeu, donc à la structure du match. L'analyse de la succession des unités logiques permet donc trouver des macro-segments de plus haut-niveau. Suite à une telle segmentation, chaque plan élémentaire de la vidéo sera assigné à un niveau de hiérarchie décrit en terme d'unités logiques, de point, de jeu et de set.

Pour atteindre cet objectif, nous modélisons l'entrelacement temporel des unités logiques par des modèles de Markov cachés hiérarchiques.



(a) (b)

FIG. 4.1: A gauche, structure intrinsèque d'un match de tennis. A droite, structure de la vidéo d'un match de tennis.

4.2 Modélisation de la structure globale

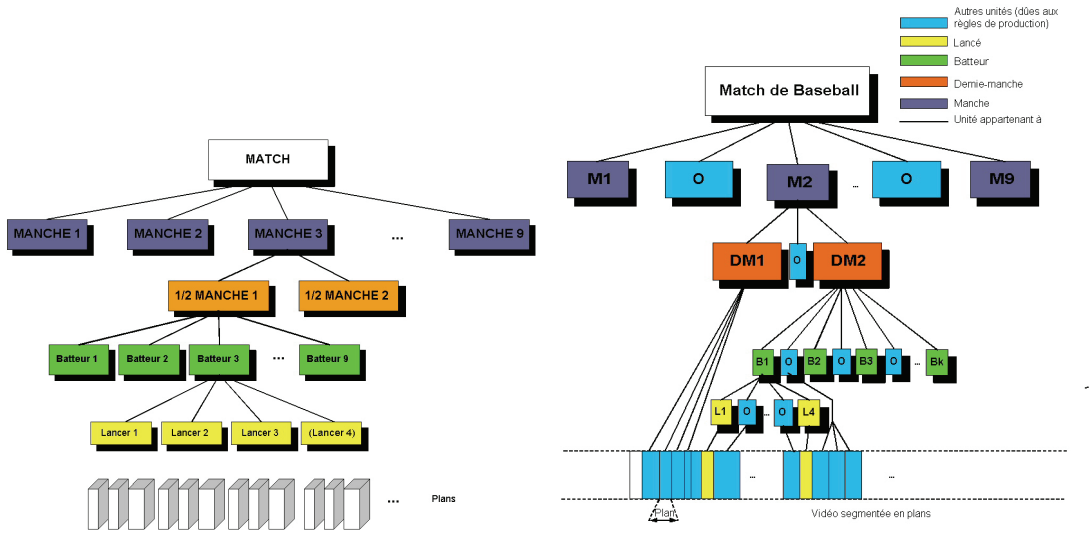
4.2.1 HMM hiérarchiques

Les HMMs hiérarchiques (HHMMs pour *Hierarchical Hidden Markov Models*) sont des processus stochastiques à plusieurs niveaux. Ils généralisent les HMMs standards en considérant chaque état caché comme un modèle probabiliste à part entière, *i.e.* chaque état est lui-même un HHMM. Un état génère donc une séquence de symboles plutôt qu'un seul symbole. Les séquences sont émises par activation des sous-états d'un état. Ces sous-états peuvent eux-mêmes être composés de sous-états qu'ils activent, etc. Ce processus récursif prend fin lorsqu'il atteint un état spécial que nous appellerons "état émetteur". Les états émetteurs sont les seuls états qui émettent réellement un symbole d'observation selon le mécanisme usuel d'émission des HMMs. Les états cachés qui ne produisent pas directement de symboles d'observation sont appelés "états internes". L'activation d'un sous-état par un état interne est appelé "transition verticale". Une transition entre états du même niveau est appelée "transition horizontale".

L'ensemble des états et des transitions verticales induit une structure en arbre dans laquelle l'état racine est le noeud au sommet de la hiérarchie, et les états émetteurs sont les feuilles. Toutes les feuilles ne sont bien entendu pas nécessairement à la même distance de la racine.

Les HHMMs permettent de corréler des événements qui se produisent loin les uns des autres dans une séquence d'observations. Ils ont été appliqués pour la reconnaissance d'écriture manuscrite et l'analyse de textes [130].

Nous allons donner une description formelle d'un HHMM. La figure 4.3 illustre la structure d'un HHMM de topologie et paramètres arbitraires. D est le nombre de niveaux de la hiérarchie. On note $d \in \{1, ..., D\}$ l'index de la hiérarchie : cet index est égal à 1 à la racine, et à D pour les états émetteurs (sauf si tous les états émetteurs ne sont pas à la même distance de la racine). Dans la figure 4.3, $D = 4$. Un HHMM est défini par un



(a) (b)

FIG. 4.2: A gauche, structure intrinsèque d'un match de baseball. A droite, structure de la vidéo d'un match de baseball.

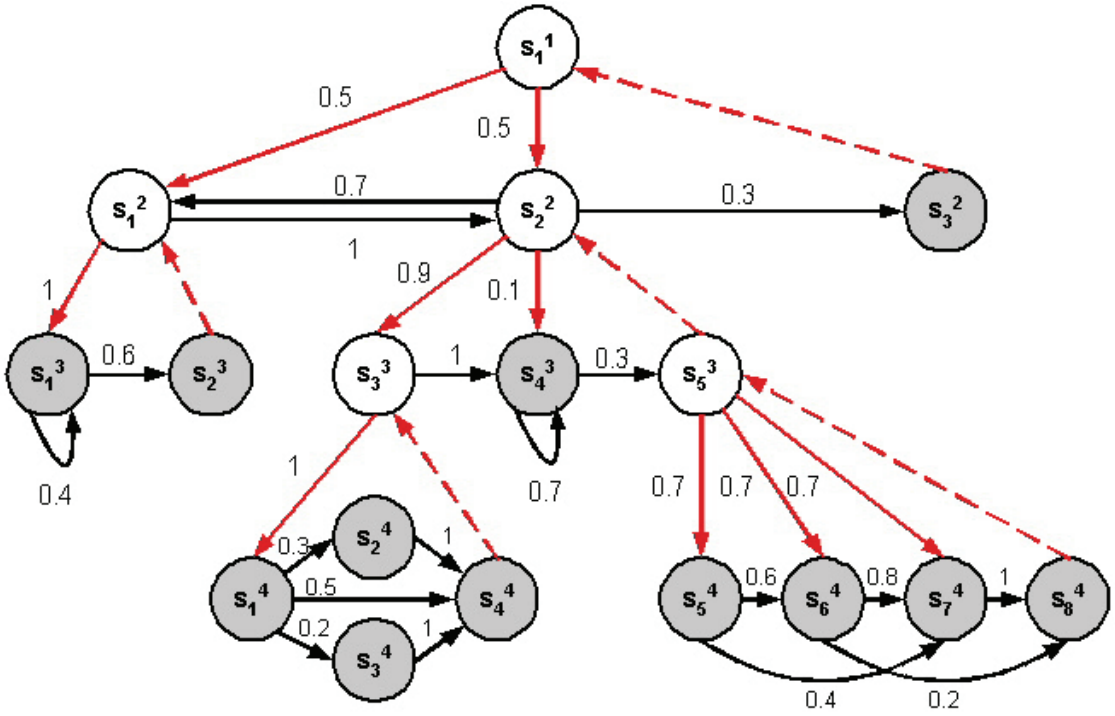


FIG. 4.3: Exemple de HHMM à 4 niveaux. Les transitions rouges représentent les transitions verticales, et les noires les transitions horizontales. Les états grisés sont les états émetteurs.

ensemble S^I d'états internes, et un ensemble S^E d'états émetteurs. L'état d'index i au niveau d de la hiérarchie se note s_i^d. Un état s_i^d peut être un état émetteur ou un état interne.

Si $s_i^d \in S^E$ est un état émetteur, il est entièrement caractérisé par son vecteur des probabilités d'observations :

$$B^{s_i^d} = (b_{s_i^d}(k))_{1 \leq k \leq M}$$

où $b_{s_i^d}(k) = P(v_k|s_i^d)$ est la probabilité d'observer le symbole v_k en étant dans l'état s_i^d. Le symbole v_k appartient à l'ensemble discret $V = \{v_1, v_2, ..., v_M\}$ des M symboles d'observation associés aux états émetteurs.

Si $s_i^d \in S^I$ est un état interne, il est caractérisé par un ensemble de $N^{s_i^d}$ sous-états notés $S^{s_i^d} = \{s_{N+1}^{d+1}, s_{N+2}^{d+1}, ..., s_{N+N^{s_i^d}}^{d+1}\}$ avec $N = \sum_{k=1}^{i_1} N^{s_k^d}$. Par exemple, dans la figure 4.3, l'état interne de niveau 2 s_2^2 possède $N^{s_2^2} = 3$ sous-états : $S^{s_2^2} = \{s_3^3, s_4^3, s_5^3\}$. Pour être entièrement déterminé, l'état s_i^d est encore caractérisé par la matrice $A^{s_i^d}$ des probabilités de transitions horizontales entre ses sous-états et par les probabilités $\Pi^{s_i^d}$ d'activation initiale de chaque sous-état. On a :

$$A^{s_i^d} = (a_{kl}^{s_i^d})_{N+1 \leq k,l \leq N+N^{s_i^d}}, \text{ avec } a_{kl}^{s_i^d} = P(s_l^{d+1}|s_k^{d+1})_{s_k^{d+1}, s_l^{d+1} \in S^{s_i^d}}$$

et

$$\Pi^{s_i^d} = (\pi_k^{s_i^d})_{N+1 \leq k \leq N+N^{s_i^d}}, \text{ avec } \pi_k^{s_i^d} = P(s_k^{d+1}|s_i^d)$$

$\pi_k^{s_i^d}$ est également interprété comme la probabilité de transition verticale entre s_i^d et s_k^{d+1}. Selon la topologie du HHMM, chaque état possède un ou plusieurs sous-états "de sorties", *i.e.* qui retournent à l'état parent l'ayant activé. On note $\upsilon_k^{s_i^d} = P(s_i^d|s_k^{d+1})$ la probabilité que l'état $s_k^{d+1} \in S^{s_i^d}$ termine la transition verticale et

$$\Upsilon^{s_i^d} = (\upsilon_k^{s_i^d})_{N+1 \leq k \leq N+N^{s_i^d}}$$

la matrice des transitions de sorties. On notera que A n'est pas stochastique en raison de la séparation des transitions horizontales et verticales.

Dans l'exemple de la figure 4.3, l'état interne s_2^2 est caractérisé par la matrice $A^{s_2^2}$:

$$A^{s_2^2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 0 \end{bmatrix}$$

et par les matrices des transitions verticales :

$$\Pi^{s_2^2} = \begin{bmatrix} 0.9 & 0.1 & 0 \end{bmatrix} \quad \Upsilon^{s_2^2} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Nous formaliserons un HHMM par $\eta = (D, \{A^{S^I}\}, \{\Pi^{S^I}\}, \{\Upsilon^{S^I}\}, B^{S^E})$.

Les trois problèmes fondamentaux évoqués pour les HMMS au paragraphe 2.2.3 restent valables pour les HHMMs. Les solutions proposées doivent cependant être modifiées pour tenir compte de la structure hiérarchique et des propriétés multi-échelles des HHMMs. En effet, pour une séquence d'observation donnée, la séquence d'états la plus probable est une structure multi-résolution d'états activés, au lieu d'une simple séquence d'indices des états les plus probablement atteints.

4.2.2 Application à la structure d'un match

L'idée est de modéliser la structure hiérarchique d'un match de tennis par un HHMM. Un match de tennis est composé de plusieurs sets. Chaque set est lui-même décomposé en plusieurs jeux, et chaque jeu en points.

Comme pour la modélisation des unités logiques, l'information *a priori* est utilisée pour déterminer le nombre d'états et la topologie du HHMM. Les informations prises en compte sont celles relatives à la structure d'un match de tennis.

Le modèle créé est représenté dans la figure 4.4. Le nombre de niveaux de hiérarchie du HHMM découle directement de celui du tennis : $D = 6$. Chaque niveau est décrit ci-dessous d'après le formalisme mis en place à la section précédente.

niveau "match" : $d = 1$
Il s'agit de la racine du HHMM. Il n'est donc composé que d'un seul état noté M. Dans le cas d'un match en deux sets gagnants, l'état M est composé de $N^M = 5$ sous-états notés : $S^M = \{S_1, TR_1, S_2, TR_2, S_3\}$. Les états S_1, S_2 et S_3 correspondent aux sets, tandis que les états $TR_1 et TR_2$ représentent les temps de repos liés à la fin d'un set, et aux éventuels changements de côté. Dans le cas d'un match en trois sets gagnants, on aurait $N^M = 9$.

L'état M est caractérisé par les matrices de transitions horizontales et verticales entre les sous-états : $A^M = \{a^M_{S_k S_l} = P(S_l|S_k)\}_{S_k,S_l \in S^M}$, $\Pi^M_{S_k} = \{P(S_k|M)_{S_k \in S^M}\}$ et $\Upsilon^M_{S_k} = \{P(M|S_k)_{S_k \in S^M}\}$.

niveau "set" : $d = 2$
Pour un match en deux sets gagnants, le niveau est donc composé de 3 états modélisant les sets notés $\{S_1, S_2, S_3\}$, et de 2 états modélisant les temps de repos notés $\{TR_1, TR_2\}$.

Les états $\{S_1, S_2, S_3\}$ possèdent 10 sous-états notés : $S^{S_i} = \{J_1, CC_1, J_2, J_3, CC_2, J_4, J_5, CC_3, J_6, J_7\}$ $1 \leq i \leq 3$. Les états $\{J_1, J_2, ..., J_7\}$ modélisent les jeux. Autant d'états de "jeu" sont nécessaires pour assurer que le HHMM passera au moins six fois dans un état "jeu". Les matrices de transitions horizontales et verticales associées à chaque état sont :

$$A^{S_i} = \{a^{S_i}_{J_k J_l} = P(J_l|J_k)\}_{J_k,J_l \in S^{S_i}},$$

$$\Pi^{S_i}_{J_k} = \{P(J_k|S_i)\}_{J_k \in S^{S_i}} \text{ et } \Upsilon^{S_i}_k = \{P(S_i|J_k)\}_{J_k \in S^{S_i}}$$

Les sous-états des TR_i sont les états émetteurs correspondant au HMM représentant l'unité logique "temps morts" définie dans la section 2.3.2. On note :

$$A^{TR_i} = \{a^{TR_i}_{s_k s_l} = P(s_l|s_k)\}_{1 \leq k,l \leq N^{TR}}$$

et

$$\Pi^{TR_i}_{s_k} = \{P(s_k|TR_i)\}_{1 \leq k \leq N^{TR}} \text{ et } \Upsilon^{TR_i}_k = \{P(TR_i|s_k)\}_{1 \leq k \leq N^{TR}}$$

niveau "jeu" : $d = 3$
Il se compose de 7 états modélisant les jeux notés $\{J_1, J_2, ..., J_7\}$ et de 3 états $\{CC_1, CC_2, CC_3\}$ modélisant les changements de côté de joueurs à la fin du premier jeu, puis tous les deux jeux.

Chaque état "jeu" se décompose en 5 sous-états $S^{J_i} = \{P_1, P_2, ..., P_5\}$ correspondant aux points. Comme pour les états TR_i du niveau supérieur, les sous-états des CC_i sont les états émetteurs correspondant au HMM de l'unité logique "temps morts".

Les matrices de transitions pour ce niveau se notent :

$$A^{J_i} = \{a^J_{P_k P_l} = P(P_l|P_k)\}_{1 \leq k,l \leq N^J},$$

$$\Pi^{J_i}_{P_k} = \{P(P_k|J_i)\}_{1 \leq k,l \leq N^J} \text{ et } \Upsilon^{S_i}_k = \{P(J_i|P_k)\}_{1 \leq k,l \leq N^J}$$

et

$$A^{CC_i} = \{a^{CC_i}_{s_k s_l} = P(s_l|s_k)\}_{1 \leq k,l \leq N^{CC}},$$

$$\Pi^{CC_i}_{s_k} = \{P(s_k|CC_i)\}_{1 \leq k \leq N^{CC}} \text{ et } \Upsilon^{CC_i}_k = \{P(CC_i|s_k)\}_{1 \leq k \leq N^{CC}}$$

niveau "point" : $d = 4$

Il se compose de 5 états notés $\{P_1, P_2, ..., P_5\}$ modélisant les points. Chaque point est composé de 3 sous-états $\{PS, E, Re\}$ qui correspondent aux unités logiques définies en 2.3.2. PS est l'unité logique "premier service manqué", E "échange" et Re "rediffusion". Un point est défini par un échange ou par un premier service manqué suivi d'un échange. Les échanges sont éventuellement suivis d'une ou de plusieurs rediffusions, qui sont ici intégrées à l'état "point".

$$A^{P_i} = \{a^{P_i}_{s_k s_l} = P(s_l|s_k)\}_{1 \leq k,l \leq N^P}$$

$$\Pi^{P_i}_{es_k} = \{P(es_k|CC_i)\}_{es_k \in \{FS,E,Re\}}$$

et

$$\Upsilon^{P_i}_k = \{P(P_i|es_k)\}_{es_k \in \{FS,E,Re\}}$$

niveau "éléments structurants" : $d = 5$

Il est composé de 3 états correspondants aux unités logiques et notés $\{FS, E, Re\}$. Chaque unité logique est composé de sous-états qui sont les états émetteurs du HHMM. On note A^{FS}, A^E et A^{Re} les matrices de transitions horizontales associées et $\Pi^{FS}, \Pi^E, \Pi^{Re}, \Upsilon^{FS}, \Upsilon^E$ et Υ^{Re} les matrices de transitions verticales.

niveau "états émetteurs" : $d = 6$

Les états émetteurs sont ceux des unités logiques. Les probabilités d'observations et l'ensemble des symboles d'observations sont les mêmes que ceux définis dans le chapitre consacré aux unités logiques.

Le HHMM représenté dans la figure 4.4 possède donc 326 états internes et 1387 états émetteurs. Afin de simplifier sa mise en œuvre et réduire le nombre des distributions d'observations (1 par état émetteur) et des matrices de transitions (1 par état interne), les motifs répétitifs du HHMM sont factorisés. Les matrices de transitions horizontales et

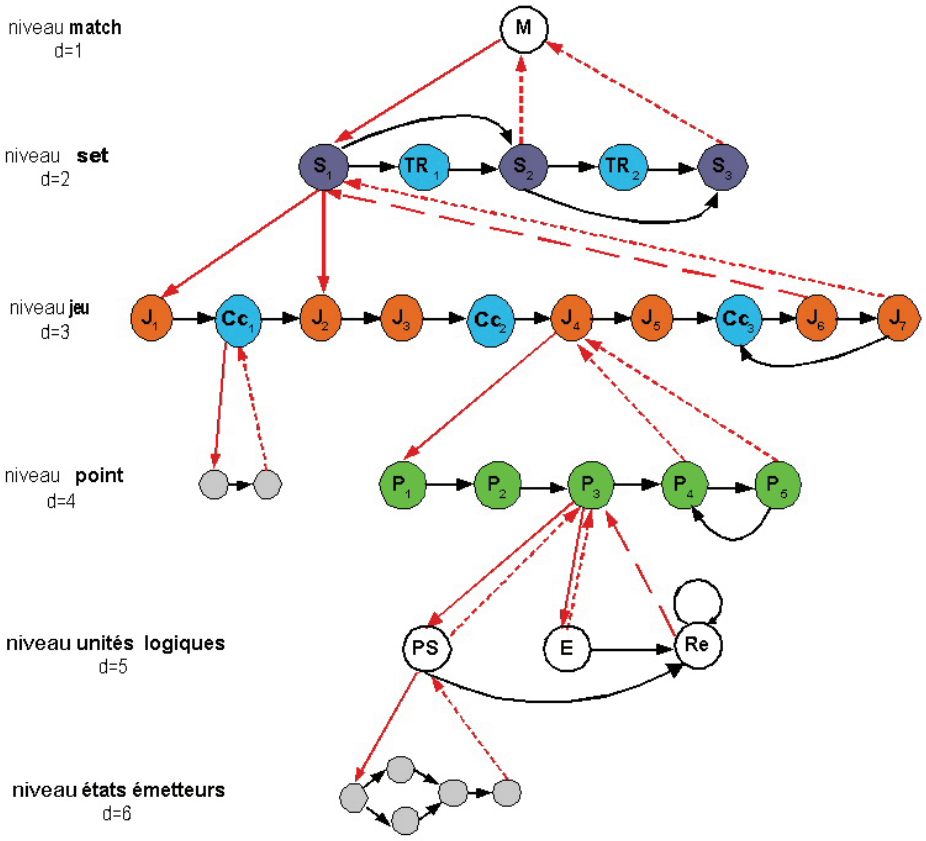


FIG. 4.4: HHMM modélisant la structure d'un match de tennis en deux sets gagnants.

verticales sont les mêmes pour chaque état d'un niveau identique :

$$\begin{array}{ll} \forall i = 1, ..., N^S & A^{S_i} = A^S \\ \forall i = 1, ..., N^J & A^{J_i} = A^J \\ \forall i = 1, ..., N^{TR} & A^{TR_i} = A^{TR} \\ \forall i = 1, ..., N^{CC} & A^{CC_i} = A^{CC} \\ \forall i = 1, ..., N^P & A^{P_i} = A^P \end{array} \tag{4.1}$$

De la même façon, les unités logiques sont identiques pour tous les états P_i, CC_i, et TR_i. Ces états partagent donc les mêmes matrices de transitions horizontales $A^{FS}, A^E, A^{Re}, A^{CC}$ et verticales $\Pi^{FS}, \Pi^E, \Pi^{Re}, \Pi_{CC}, \Upsilon^{FS}, \Upsilon^E, \Upsilon^{CC}$ et Υ^{CC}.

Après factorisation, le HHMM possède toujours le même nombre d'états. Cependant, il n'a plus que 8 matrices de transitions horizontales, dont 4 correspondent aux matrices de transitions des unités logiques.

On procède de la même façon pour le baseball et on définit le HHMM représenté dans la figure 4.5.

4.2.3 Apprentissage et décodage

Les matrices de transitions horizontales et verticales des unités logiques sont estimées dans une phase d'apprentissage identique à celle présentée au chapitre 2, car elles sont étroitement liées au montage vidéo. En revanche, les autres matrices (A^M, A^S, A^J, A^P,...) relèvent uniquement des règles du tennis. En ce sens, elles ne peuvent être estimées, mais doivent être fixées *a priori*. En effet, la probabilité qu'un set soit composé de 9 jeux plutôt que 8 ou 6 n'a pas de sens. Ce sont les observations, donc le contenu de la vidéo, qui vont déterminer les transitions au niveau de la structure.

Les matrices horizontales A^M, A^S, A^J, A^P et verticales $\Pi^M, \Pi^S, \Pi^J, \Pi_P, \Upsilon^M, \Upsilon^S, \Upsilon^J$ et Υ^P sont fixées de la manière suivante :

$$A^M = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \Pi^M = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \Upsilon^M = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0.5 \end{bmatrix} \tag{4.2}$$

$$A^S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \end{bmatrix} \quad \Pi^{S_i} = \begin{bmatrix} 0.5 \\ 0 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \Upsilon^{S_i} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0.5 \end{bmatrix} \tag{4.3}$$

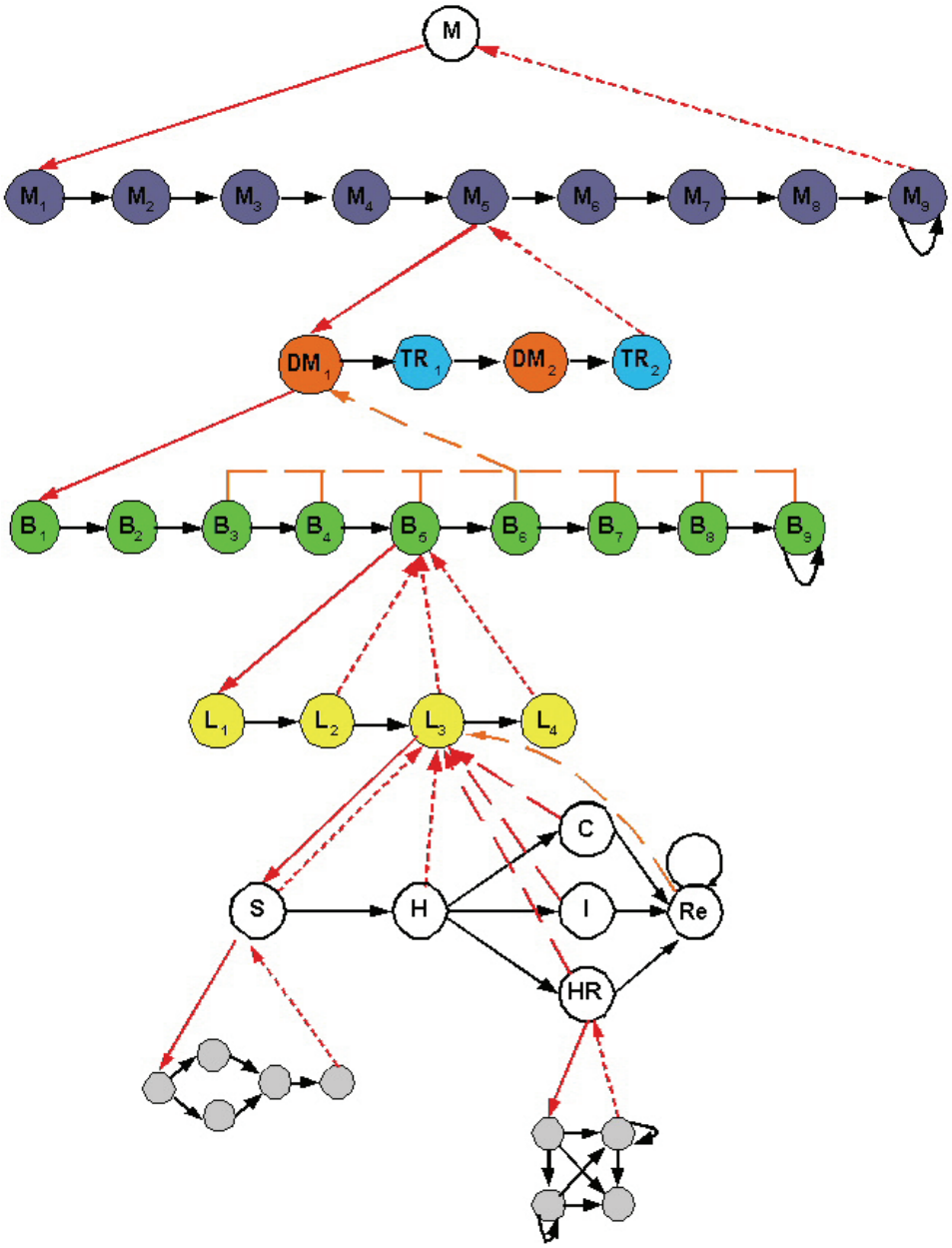


FIG. 4.5: HHMM modélisant la structure d'un match de baseball.

$$A^J = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 \end{bmatrix} \quad \pi^{J_i} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \upsilon^{J_i} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.5 \\ 0.5 \end{bmatrix} \tag{4.4}$$

Le décodage est réalisé par un algorithme de Viterbi généralisé aux HHMMs, qui cherche la séquence d'états la plus probable dans l'ensemble des états du HHMM. Le HHMM tel qu'il a été défini pose cependant un problème lors du décodage, lié à la définition a priori des probabilités de transitions verticales équiprobables entre les niveaux. Prenons l'exemple d'une transition possible entre deux jeux J_2 et J_3. Supposons qu'à l'instant t, l'observation o_t soit dans l'état émetteur de sortie s_i de l'état P_5 du jeu J_2. Nous noterons cet état i sous la forme du vecteur $[s_i\ P_5\ J_2]$. A l'instant $t+1$, d'après les probabilités de transitions, le système soit reste dans l'état J_2, soit passe dans l'état J_3. Depuis l'état i, l'algorithme de Viterbi va calculer à $t+1$ quel état j est le plus vraisemblable par la quantité : $\delta_{t+1}(j) = a_{ij} b_j(o_{t+1})$. Pour l'état $j' = [s_j P_4 J_2]$, $\delta_{t+1}(j')$ s'écrit :

$$\delta_{t+1}(j') = p(P_4|P_5)\ b_j(o_{t+1})$$

Pour l'état $j" = [s_j P_1 J_3]$, $\delta_{t+1}(j")$ s'écrit :

$$\delta_{t+1}(j") = p(P_5|J_3)\ p(J_3|J_2)\ p(J_2|P_1)\ b_j(o_{t+1})$$

Pour un même état émetteur s_j, les probabilités d'observations $b_j(o_{t+1})$ sont identiques pour tous les états P_i. De même on a défini : $p(P_4|P_5) = p(P_5|J_3)$ et $p(J_3|J_2) = p(J_2|P_1) = 1$. Il en résulte que $\delta_{t+1}(j') = \delta_{t+1}(j")$.

Les probabilités de transitions entre deux états possibles ne favorisent pas une transition plutôt qu'une autre et c'est finalement le sens de parcours des états par l'algorithme de Viterbi qui va déterminer la transition. Ces probabilités sont fixées *a priori* et les modifier n'aurait aucun sens. D'autre part, les probabilités d'observations liées à la similarité visuelle, à la durée du plan et aux vecteurs audio sont indépendantes des états internes du HHMM.

Il faut donc déterminer un nouvel attribut dont la probabilité d'observation pourrait guider le processus de décodage. Nous allons voir dans la section suivante en quoi cet attribut peut être la position du joueur et comment intégrer cette information supplémentaire.

4.3 Extraction d'attributs spécifiques

La position du joueur par rapport au terrain au moment du service permet de déterminer quel joueur engage le point. L'identification du serveur est porteuse d'un certain nombre d'informations sur l'état du match. Les règles suivantes spécifiques au service sont exploitées dans la suite pour extraire ces informations :

- en exécutant le service, le serveur doit se tenir alternativement derrière la moitié droite et la moitié gauche du court en commençant à droite dans chaque jeu ;
- à la fin du premier jeu, le relanceur devient serveur et le serveur relanceur, et ainsi de suite, alternativement, pour tous les jeux d'une partie ;

– les joueurs doivent changer de côté à la fin du 1er jeu, puis tous les 2 jeux.

La figure 4.6 résume la position du joueur en bas de l'image en fonction du déroulement du jeu, d'après les règles énoncées ci-dessus. Cette figure illustre bien qu'un changement de point ou de jeu peut être déduite de la position du serveur.

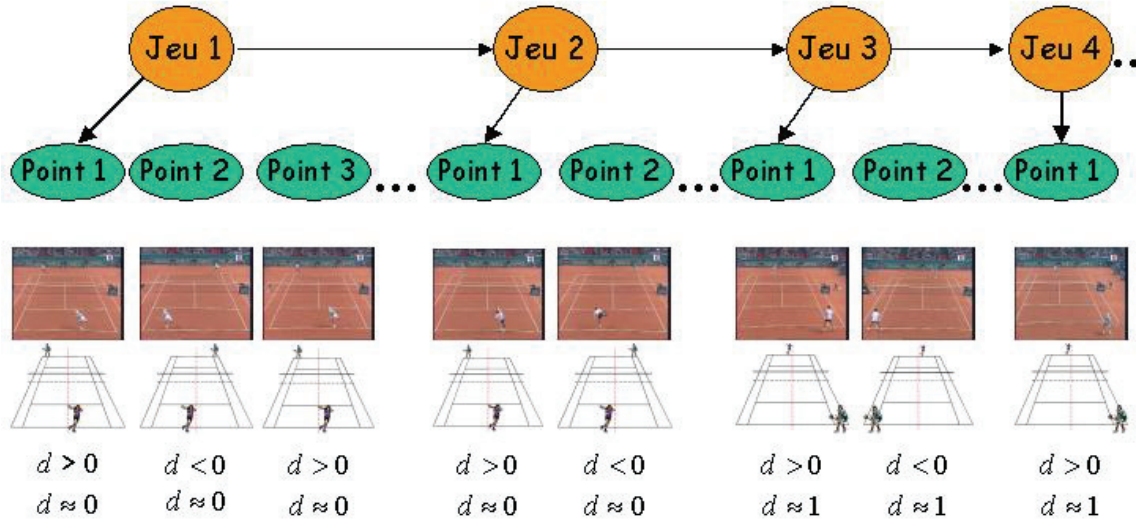
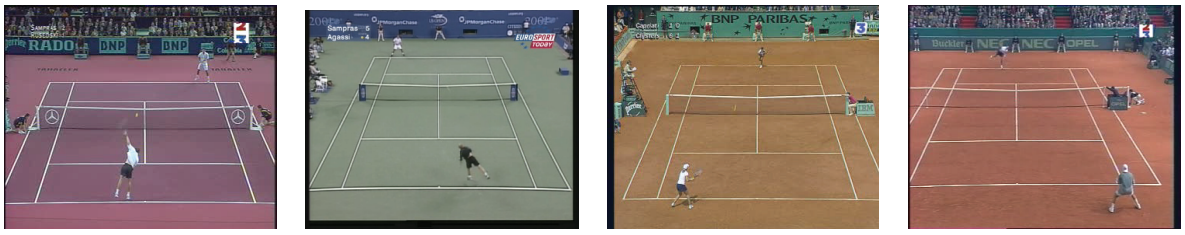


FIG. 4.6: Evolution de la position du joueur du bas de l'image au service au cours d'un set.

Il faut encore détecter et localiser les joueurs afin de déterminer quel joueur engage le point. Le serveur est identifié en déterminant la position du joueur par rapport au terrain au début du jeu. Généralement, le joueur ayant le service est situé près de la ligne centrale de service, tandis que le joueur qui réceptionne est excentré du terrain, proche des couloirs (Fig. 4.7). L'indice que nous avons choisi de calculer afin de déterminer le serveur est donc la distance du joueur par rapport à la ligne centrale de service.

Nous allons d'abord décrire la méthode utilisée pour calculer cette distance : détection et recalage du terrain, détection du joueur, et enfin calcul de la distance distance du joueur par rapport à la ligne centrale de service. Nous expliquerons ensuite comment la traduire en terme de probabilités d'observations.



(a) (b) (c) (d)

FIG. 4.7: Position des joueurs en début de jeu. (a)-(b) le joueur du bas sert, (c)-(d) le joueur du bas réceptionne.

4.3.1 Recalage du terrain

Nous avons choisi d'utiliser le modèle théorique du terrain de tennis (dont les dimensions sont parfaitement connues) afin d'identifier l'ensemble du terrain sur une image provenant d'une séquence de jeu. Le problème de la détection et de la reconnaissance des lignes du terrain se transforme alors en un problème d'identification du modèle de déformation permettant de déformer le terrain théorique caractérisé par des angles droits, en un terrain vu au travers d'une caméra, où les lignes théoriques verticales recalées ne sont plus parallèles (Fig. 4.8).

Modèle de déformation

Le type de projection sous-jacente lié à un modèle classique de caméra est le modèle de projection perspective (les déformations non linéaires liées aux imperfections optiques de la caméra sont négligées). Sous l'hypothèse de projection perspective pure (sans déformation non linéaire), il existe un modèle exact permettant de transformer un plan en sa projection. Ce modèle est le modèle homographique à huit paramètres. Les images (x', y') du point (x, y) par une telle transformation sont de la forme :

$$\begin{aligned} x'_i &= \frac{h_0 x_i + h_1 y_i + h_2}{h_6 x_i + h_7 y_i + 1} \\ y'_i &= \frac{h_3 x_i + h_4 y_i + h_5}{h_6 x_i + h_7 y_i + 1} \end{aligned} \tag{4.5}$$

Bien que non linéaire en coordonnées cartésiennes, le passage en coordonnées homogènes permet de retrouver une linéarité entre un point du modèle et son projeté dans l'image. Soit $\tilde{p}(x, y, t)$, un point 2D exprimé en coordonnées homogènes. Ce même point exprimé dans l'espace cartésien aura pour coordonnées $p(\frac{x}{t}, \frac{y}{t})$. En coordonnées homogènes, la transformation homographique s'exprime sous forme matricielle via la matrice $H(3 \times 3)$ définie à un coefficient multiplicatif λ près :

$$A^M \propto \begin{bmatrix} h_0 & h_1 & h_2 \\ h_3 & h_4 & h_5 \\ h_6 & h_7 & 1 \end{bmatrix} \tag{4.6}$$

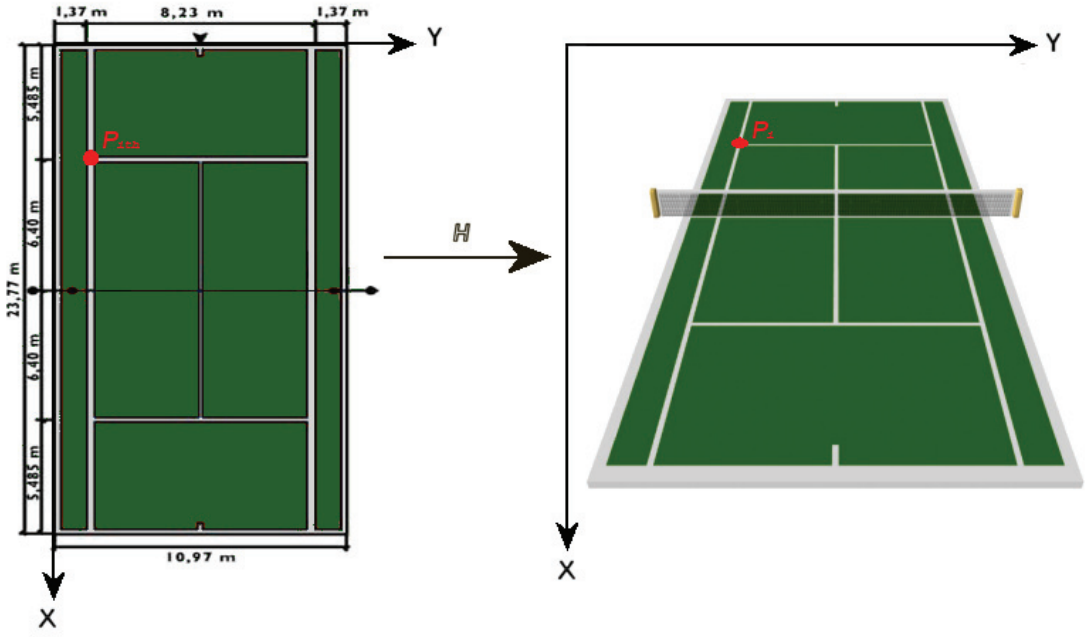


FIG. 4.8: Recalage du modèle théorique du terrain de tennis sur une image de la vidéo.

Quel que soit le point $\tilde{p}^i_{th}$ exprimé dans le repère attaché au modèle théorique, et son correspondant $\tilde{p}^i$ dans l'image, on a :

$$\tilde{p}^i = \lambda \cdot H \cdot \tilde{p}^i_{th} \tag{4.7}$$

avec λ scalaire non nul. L'équation (4.7) peut s'écrire sous la forme d'un système $AX = B$ où le vecteur inconnu $X = [h_0 h_1 h_2 h_3 h_4 h_5 h_6 h_7]^T$ est à 8 dimensions. Au minimum 4 points vérifiant l'équation (4.7) sont donc nécessaires pour calculer les 8 paramètres de l'homographie.

Dans notre cas, nous cherchons l'homographie H liant les lignes du terrain de l'image réelle à celles du modèle. Afin de résoudre le système précédent, nous devons donc identifier au moins 4 lignes de l'image.

Procédé de recalage

Les hypothèses posées sont les suivantes :
- la surface de jeu et donc la couleur du terrain n'est pas connue ;
- les lignes de terrain sont blanches ;
- le recalage s'applique aux prises de vue pour lesquelles la majeure partie du terrain est visible.

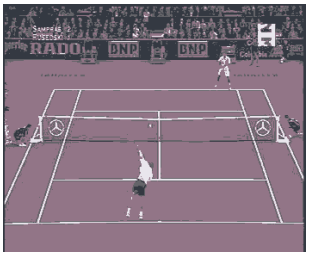
La méthode que nous proposons ne fait pas appel à une initialisation manuelle de l'algorithme de recalage pour chaque séquence traitée. Elle est par ailleurs robuste au problème de détection de contours. La technique utilisée comporte les étapes suivantes (Fig. 4.9) :

1. extraction de l'image des couleurs dominantes ;
2. détection des contours par un filtre de Cany-Deriche pour obtenir les cartes des gradients verticaux et horizontaux ;
3. binarisation et filtrage de ces cartes par utilisation de l'information couleur (les points appartenant aux lignes étant supposés blancs) : seuls les points de fort contraste possédant un niveau de gris supérieur à un certain seuil sont conservés ;
4. transformation de Hough afin de détecter les lignes parmi les contours candidats et identification des lignes (Fig. 4.10) ;
5. enfin, calcul de l'homographie H recalant le terrain théorique sur l'image réelle, à l'aide d'au moins 4 des lignes identifiées.

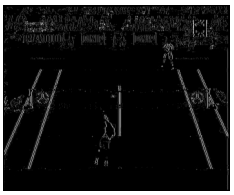
Qualité du recalage

La phase de recalage consiste à identifier l'homographie recalant le terrain théorique sur l'image réelle. Comme le montre la figure 4.8, la déformation à appliquer au modèle théorique est très importante. De par sa nature fortement non linéaire en coordonnées cartésiennes, le modèle homographique est relativement instable (des petites variations sur les paramètres de la troisième ligne de la matrice entraînent de très fortes variations de la position des points recalés).

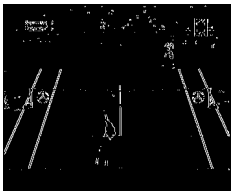
Le critère de qualité que nous nous imposons repose sur la distance moyenne entre lignes recalées et lignes réelles. Ce critère $C(I, H)$, dépendant de l'image I et de l'homographie



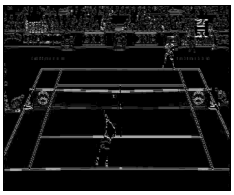
(a) image originale



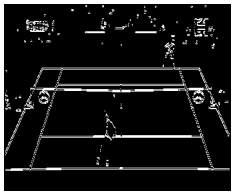
(b) Gradients horizontaux



(c) Binarisation des gradients horizontaux

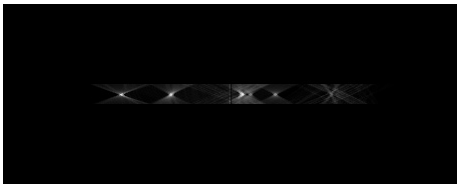


(d) Gradients verticaux

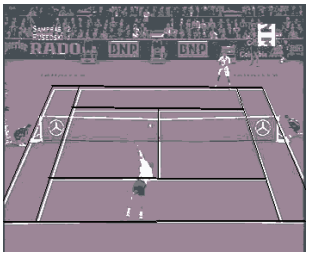


(e) Binarisation des gradients verticaux

(f) Transformée de Hough horizontale



(g) Transformée de Hough verticale



(h) Calcul de l'homographie

FIG. 4.9: Processus de recalage du modèle du terrain de tennis.

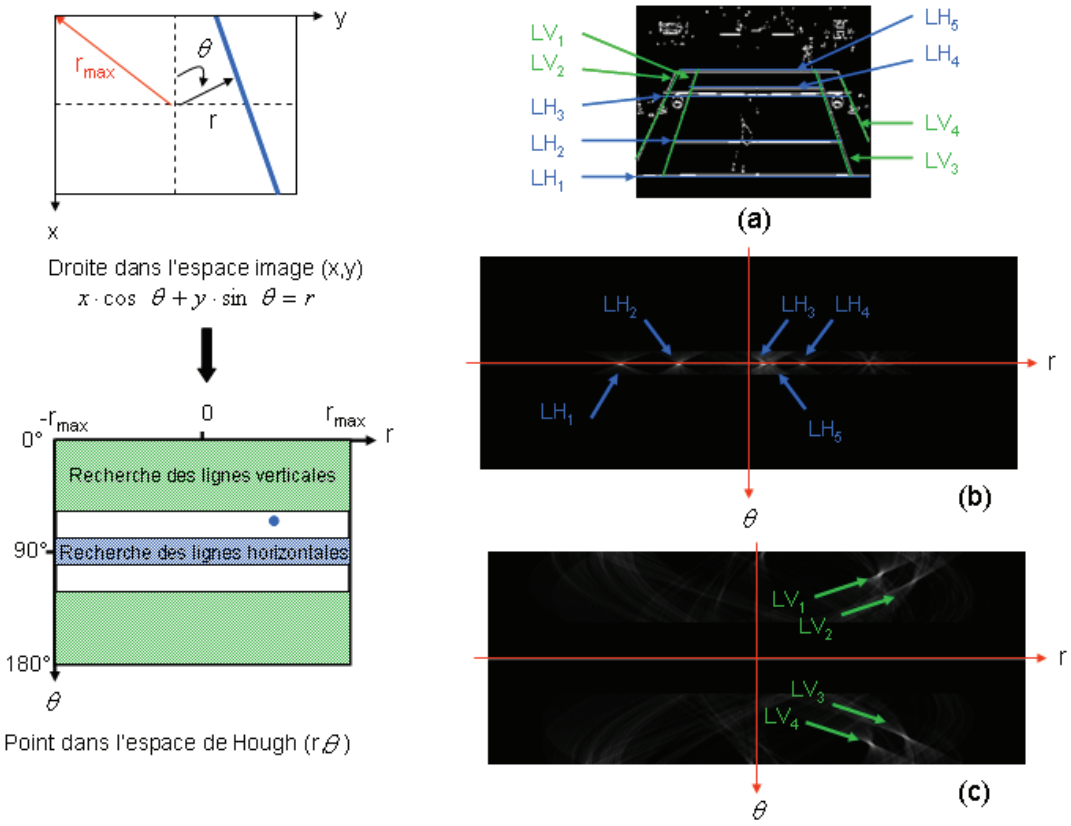


FIG. 4.10: Identification des lignes du terrain dans l'espace de Hough (r, θ).

H, est défini comme l'intégrale le long du contour du terrain T, de la distance entre un point du terrain recalé et le point de contour le plus proche. Son expression est la suivante :

$$C(I,H) = \oint_T d_c(I, H.s)ds \tag{4.8}$$

où d_c est la distance euclidienne du point s recalé par l'homographie H au point contour le plus proche dans l'image I.

Lorsque $C(I,H)$ est supérieur à un certain seuil Th_{rec}, on considère que la détection, l'identification des lignes et le calcul de l'homographie ont échoués. Dans notre implémentation, nous utilisons $Th_{rec} = 50$.

4.3.2 Positions des joueurs

L'hypothèse posée est la suivante : le positionnement des joueurs dans les images-clés est celui du service. Pour satisfaire au mieux cette hypothèse, on a pris soin de choisir les images-clés en début de plan lors du processus d'extraction des images-clés. Du fait de la petite taille des joueurs dans l'image, seul le joueur situé en bas de l'image est détecté, d'une part car sa détection sera plus fiable, et d'autre part car l'information contenue dans sa localisation est suffisante, la position de son adversaire pouvant s'en déduire.

Le calcul de la position du joueur s'effectue en 2 étapes :

1. extraction grossière du joueur (détection) ;
2. calcul de la distance à la ligne centrale de service (localisation).

Détection du joueur

(a) Image des couleurs dominantes

(b) Filtrage et dilatation sur la couleur dominante

(c) Détection du joueur

FIG. 4.11: Processus de detection du joueur.

La méthode utilisée comporte les étapes suivantes (Fig. 4.11) :

1. extraction de l'image des couleurs dominantes ;
2. filtrage sur la couleur dominante c_{dom} afin de séparer la zone de jeu : une opération de dilatation avec un élément structurant carré est utilisée pour supprimer les lignes du terrain dans la zone de jeu ;

3. détection grossière du joueur situé en bas de l'image. La détection s'effectue simplement à l'aide d'une fenêtre coulissante W dont la taille est le dixième de la hauteur de l'image, en cherchant la position $(\hat{x}_j, \hat{y}_j)$ du centre de la fenêtre pour laquelle :

$$(\hat{x}_j, \hat{y}_j) = \max_{x,y \in I_{inf}} \sum_{p \in W} \Omega(p) \tag{4.9}$$

où I_{inf} est la moitié inférieure de l'image, p est un pixel de la fenêtre W, et :

$$\Omega(p) = \begin{cases} 0 & \text{si la couleur du pixel p est } c_{dom} \\ 1 & \text{sinon} \end{cases}$$

Localisation du joueur

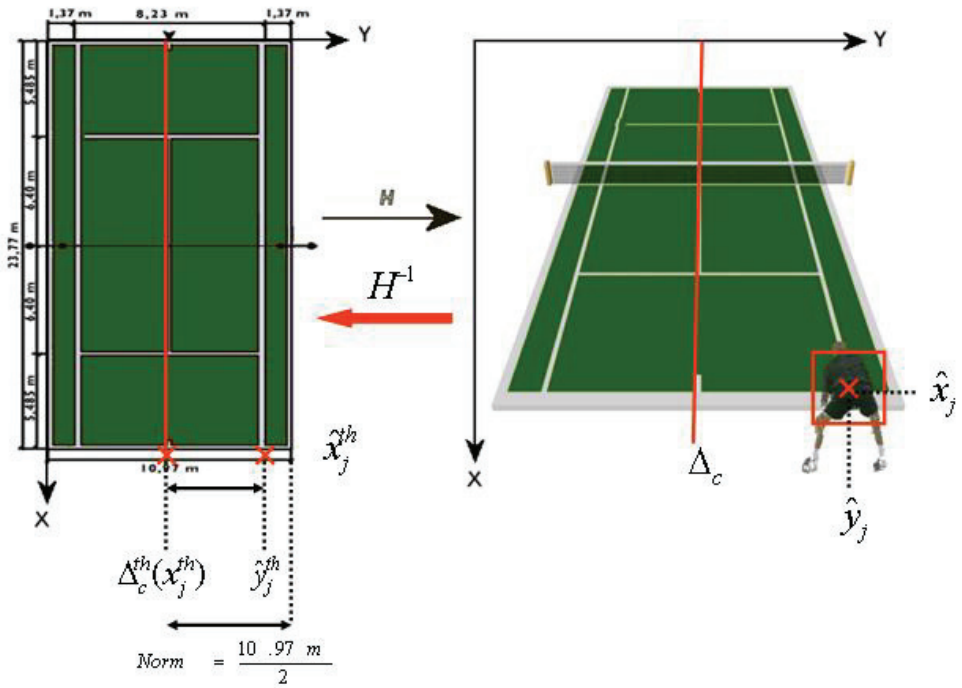


FIG. 4.12: Calcul de la distance du joueur à la ligne centrale de service.

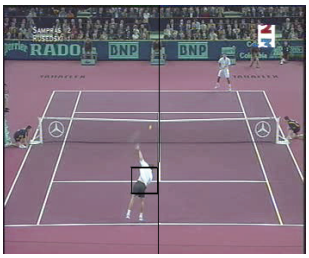
Une fois le joueur détecté, sa distance par rapport à la ligne centrale de service est calculée. La position considérée du joueur est $(\hat{x}_j, \hat{y}_j)$. L'équation de la ligne centrale $\triangle_c$ est directement calculée à partir du recalage du modèle théorique du terrain, expliqué au paragraphe précédent. La distance $d((\hat{x}_j, \hat{y}_j), \triangle_c)$ est calculée dans l'espace recalé et normalisée par rapport à la largeur théorique du terrain, comme l'indique la figure 4.12.

Cela permet d'obtenir une mesure de la distance indépendante de la prise de vue de l'image. On définit :

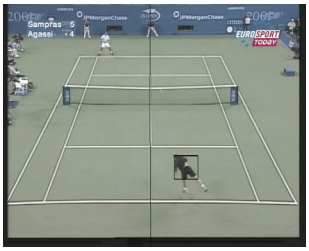
$$d((\hat{x}_j, \hat{y}_j), \triangle_c) = \begin{cases} \frac{\hat{y}^j_{th} - \triangle^c_{th}(\hat{x}^j_{th})}{\text{Norm}} & > 0 \quad \text{si le joueur est à droite de la ligne centrale} \\ \frac{\hat{y}^j_{th} - \triangle^c_{th}(\hat{x}^j_{th})}{\text{Norm}} & < 0 \quad \text{si le joueur est à gauche de la ligne centrale} \end{cases}$$

avec $\text{Norm} = \frac{\text{largeur théorique du terrain}}{2}$.

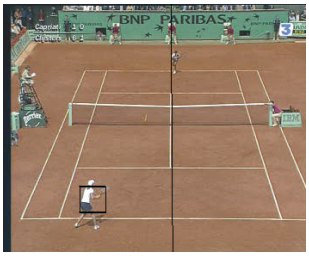
La figure 4.13 illustre les distances obtenues pour des images issues de différentes vidéos.



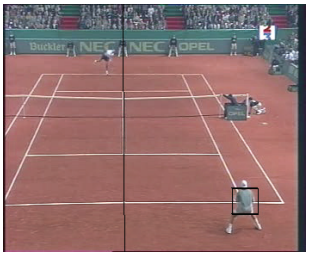
(a) $d((\hat{x}_j, \hat{y}_j), \triangle_c) = -0.09$



(b) $d((\hat{x}_j, \hat{y}_j), \triangle_c) = 0.28$



(c) $d((\hat{x}_j, \hat{y}_j), \triangle_c) = -0.55$



(d) $d((\hat{x}_j, \hat{y}_j), \triangle_c) = 0.77$

FIG. 4.13: Position détectée des joueurs en début de jeu.

4.3.3 Probabilité d'observation liée à la position du joueur

Comme nous n'avons pas différencié et reconnu les joueurs (joueur 1 en bas, au service, joueur 2 en haut, à la réception), l'information de distance à la ligne à l'instant d_t n'est pas exploitable directement, mais relativement à la position de la phase de jeu précédente d_p. Des règles énoncées sur la position du joueur au service, nous pouvons distinguer les 4 cas de figure suivants :

le joueur change de côté par rapport à la ligne centrale de service : entre chaque point. Cela se traduit par :

$$d_t \approx d_p$$

$$signe(d_t) \neq signe(d_p)$$

le joueur ne change pas de côté : après un premier service raté

$$d_t \approx d_p$$

$$signe(d_t) = signe(d_p)$$

le serveur change : entre chaque jeu pair et impair

$$d_t \ll d_p \text{ ou } d_t \gg d_p$$

$$signe(d_t) > 0$$

les joueurs changent de côté de terrain et le serveur change : entre chaque jeu impair et pair

$$d_t \approx d_p$$

$$signe(d_t) > 0$$

Chaque état j émet un symbole d'observation d_t tel que $-1 \leq d_t \leq 1$ avec une probabilité $p(d_t|j, d_p)$ dont la loi dépend de la précédente position du joueur. Il faut souligner à ce stade que l'hypothèse de l'indépendance conditionnelle des observations par rapport aux états, et des états par rapport aux états précédents reste vraie.

A l'instant t, une observation $o_t = d_t$ est bien indépendante des états $q_1, ..., q_{t-1}$ puisque c'est l'observation $o_{t-k} = d_p$ qui est prise en compte dans la loi de probabilité et non l'état q_{t-k} dans laquelle elle se trouve. En revanche o_t dépend de o_{t-k} :

$$\begin{aligned} P(o_1, ..., o_t|q_1, q_2, ..., q_t) &= P(o_t|o_1, ..., o_{t-1}, q_1, q_2, ..., q_t)\ P(o_1, ..., o_{t-1}|q_1, q_2, ..., q_{t-1}) \\ &= P(o_t|o_1, ..., o_{t-1}, q_t)\ P(o_1, ..., o_{t-1}|q_1, q_2, ..., q_{t-1}) \end{aligned} \quad (4.10)$$

En ce qui concerne l'algorithme de Viterbi, on a à présent :

$$\hat{Q} = \arg\max_Q \prod_t p(q_t|q_{t-1}) \prod_t p(o_t|q_t, o_{t-k}) \quad (4.11)$$

au lieu de :

$$\hat{Q} = \arg\max_Q \prod_t p(q_t|q_{t-1}) \prod_t p(o_t|q_t) \quad (4.12)$$

La vraisemblance $p(d_t|j, d_p)$ associée à d_t varie selon que l'état j considéré représente l'un des cas suivants :

N : la position des joueurs ne doit pas être considérée

$$p(d_t|j, d_p) = \begin{cases} 1 - \varepsilon & \text{si } d_t = 0 \\ \varepsilon & \text{si } d_t \neq 0 \end{cases}$$

CC : le joueur change de côté par rapport à la ligne centrale de service

$$p(d_t|j, d_p) = \begin{cases} 1 - ||d_t| - |d_p|| & \text{si signe}(d_t) \neq \text{signe}(d_p) \\ \varepsilon & \text{si } d_t = 0 \text{ ou signe}(d_t) = \text{signe}(d_p) \end{cases}$$

NCC : les joueurs restent à leur place

$$p(d_t|j, d_p) = \begin{cases} 1 - ||d_t| - |d_p|| & \text{si signe}(d_t) = \text{signe}(d_p) \\ \varepsilon & \text{si } d_t = 0 \text{ ou signe}(d_t) \neq \text{signe}(d_p) \end{cases}$$

SC : le serveur change

$$p(d_t|j, d_p) = \begin{cases} ||d_t| - |d_p|| & \text{si } d_t > 0 \\ \varepsilon & \text{si } d_t \leq 0 \end{cases}$$

SCC : le serveur change et les joueurs changent de côté

$$p(d_t|j, d_p) = \begin{cases} 1 - ||d_t| - |d_p|| & \text{si } d_t > 0 \\ \varepsilon & \text{si } d_t \leq 0 \end{cases}$$

4.4 Résultats expérimentaux

4.4.1 Détection et localisation du joueur

Les résultats sur le recalage du terrain et l'extraction des joueurs sont présentés dans la table 4.1. Dans notre processus, nous voulons éviter les fausses détections. En plus de la mesure de confiance sur le recalage du terrain, une contrainte sur la similarité visuelle est ajoutée. Le seuil utilisé n'a pas besoin d'être précis puisqu'il ne s'agit pas de faire une classification en vues du terrain ou non, mais bien d'éviter de réaliser des estimations coûteuses sur des plans qui sont très éloignés des vues du terrain.

Le tableau 4.1 indique les résultats de la classification des joueurs sur les séquences que nous utilisons dans cette partie. Il indique si un joueur a été détecté mais pas si la localisation est correcte.

Vidéos	*précision*	*rappel*
RG01_set1	100	92
DavisCup_set1	99	99
DavisCup_set2	100	97
DavisCup_set3	98	98
OpenParis_set1	100	84
moyenne	99	94

TAB. 4.1: Résultats de la détection des joueurs.

4.4.2 Segmentation de la structure complète

Nous ne disposons pas d'un nombre suffisant de données pour expérimenter la structuration hiérarchique pour un match complet. Nous menons donc le décodage sur un seul set. Le modèle de Markov caché hiérarchique est représenté à la figure 4.14. L'utilisation du HHMM augmente le temps de calcul de la segmentation, car l'espace de recherche est beaucoup plus grand.

4.4.2.1 Influence de la détection du joueur

Le tableau 4.2 montre les performances de la segmentation hiérarchique visuelle selon que la position du joueur soit prise en compte ou non. On remarquera que lorsque la position du joueur n'est pas prise en compte, les performances de la classification ne sont que légèrement meilleures à celles de la segmentation visuelle en unité logique présentée au

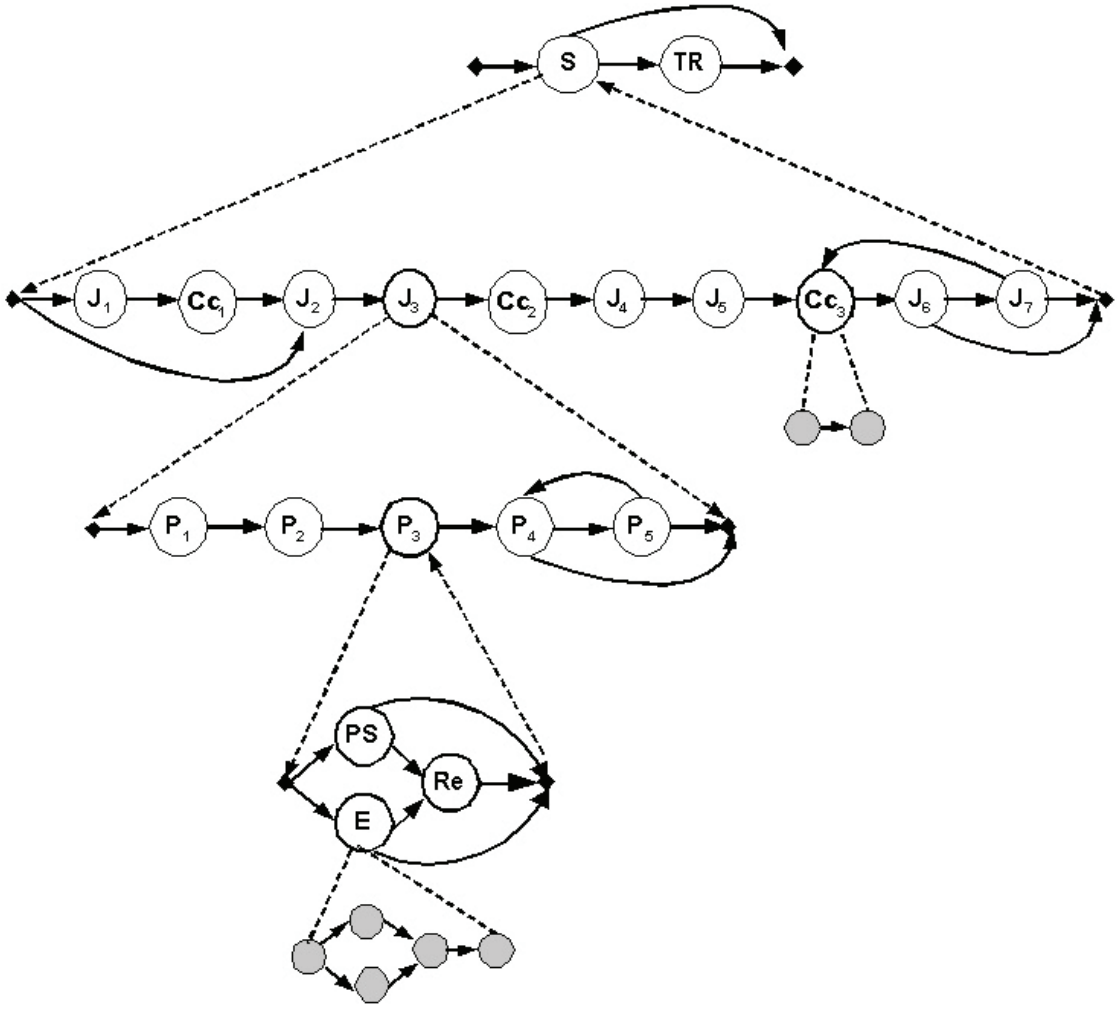


FIG. 4.14: HHMM de la structuration d'un set : 109 états internes et 463 états émetteurs.

chapitre 2. Cela montre que les probabilités de transitions estimées pour la segmentation en unités logiques ont inféré la structure de plus haut niveau du set.

Lorque la position du joueur est considérée, la segmentation est bien meilleure. Cela révèle la pertinence de cet attribut tant pour guider la structuration hiérarchique, que pour identifier les unités logiques. Il vient renforcer la similarité visuelle. Mais surtout, il lève l'ambiguïté sur les premiers services, comme le montre les résultats de la classification en unités logiques du tableau 4.3 pour la séquence RG01_set1. En effet, le joueur ne change pas de côté par rapport à la ligne centrale du terrain entre un premier service manqué et l'échange qui le suit.

	sans joueur	avec joueur
DavisCup_set1	63	84
DavisCup_set2	62	75
DavisCup_set3	58	78
OpenParis_set1	59	60
RG01_set1	65	87
moyenne	61	77

TAB. 4.2: Taux de classification globaux avec et sans la détection du joueur.

Précision	87	
Classification	Précision	Rappel
1^{er} service	89	84
Echange	92	94
Rediffusion	71	82
Temps mort	92	85
moyenne	86	86

TAB. 4.3: Classification des unités logiques avec la détection du joueur pour la séquence RG01_set1. Le taux de classification globale est de 87%.

Cependant la détection du joueur ne remplace pas la similarité visuelle. Il existe en effet des états correspondants aux vues globales pour lesquels la position du joueur n'est pas estimée. Il s'agit des plans rediffusés ou des vues du terrain pendant lesquelles les joueurs ne jouent pas.

Pour mesurer l'impact de la qualité de la détection du joueur, la segmentation est réalisée sur une séquence pour laquelle les mauvaises positions du joueur ont été corrigées. On obtient un taux de classification globale de 94%, contre un taux de 87% avec des données estimées. La loi de probabilités d'observation associée à la position du joueur dépendant de sa position précédente, le système est très sensible à la qualité de la détection de la position. Un système robuste d'extraction de la position permettrait d'atteindre des taux de classification tout à fait remarquables.

4.4.2.2 Segmentation hiérarchique audiovisuelle

La mise en œuvre des modèles de Markov cachés hiérarchiques est intéressante car elle décrit le document en terme de points et de jeux. Les frontières des points sont étroitement

liées à la segmentation des unités logiques. Les frontières entre les jeux sont uniquement déterminées par les contraintes temporelles sur le déroulement de la partie et guidées par le changement de position du joueur.

La précision de la segmentation des points et des jeux est indiqué à la table 4.4. Il suffit que les frontières soient légérement décalées par une unité logique mal identifiée, et la précision chute. Ce sont évidemment les frontières des jeux entre deux changements de côté qui sont difficiles à localiser et qui nécessitent un indice sur la position du joueur. En général, la segmentation se recale sur les changements de côté, et d'autant plus facilement si celle-ci est caractérisée par des publicités.

	points	jeux
DavisCup_set1	100	84
DavisCup_set2	97	70
DavisCup_set3	96	77
OpenParis_set1	83	30
RG01_set1	88	40
moyenne	93	60

TAB. 4.4: Précision de la segmentation en point et en jeux.

4.5 Conclusion

La segmentation en unités logiques ne permet pas d'accéder directement à la structure hiérarchique des sports étudiés. Nous avons proposé de modéliser et de segmenter l'ensemble de cette structure hiérarchique par des modèles de Markov cachés hiérarchiques. Ceux-ci sont une généralisation des modèles de Markov avec une structure de contrôle hiérarchique. L'ensemble des règles et des informations *a priori* que le système doit inférer devient alors encore plus complexe. Les informations audio-visuelles peu spécifiques que nous utilisions jusqu'alors se révèlent insuffisantes. Nous intégrons un indice de haut-niveau, spécifique au tennis, qui permet à l'algorithme de programmation dynamique de retrouver son chemin dans le graphe d'états. Il s'agit de la position du joueur au moment du service par rapport à la ligne centrale du terrain. L'alternance de la position du joueur au cours du temps fournit des indications sur le statut du jeu. Le processus de détection des joueurs par filtrage est simple et efficace mais coûteux en temps de calcul.

Le travail présenté dans ce chapitre a fait l'objet de deux publications [131, 132].

Chapitre 5

Perspectives : Représentation par modèles de segments

L'approche que nous avons proposée est basée sur le découpage de la vidéo en plans et sur leur représentation à l'aide d'attributs moyen-niveau. Cette approche présente plusieurs avantages parmi lesquels on peut citer : la rapidité et l'efficacité de l'algorithme de décodage, une bonne compréhension du processus et des lois de probabilité, et la prise en compte des incertitudes liées à l'extraction des attributs. Ces avantages ont aussi leur revers : le processus de structuration est sensible aux erreurs de segmentation temporelle, et le passage des attributs bas-niveau aux attributs moyen-niveau nécessite un traitement en plusieurs étapes.

A l'issu de cette modélisation, nous nous sommes donc interrogés sur la possibilité d'obtenir une structuration d'aussi haut-niveau à partir des attributs bas-niveau, sans réaliser aucune décision intermédiaire. Une telle approche augmenterait considérablement le temps de calcul et la complexité du processus de structuration, mais ne nécessiterait qu'une seule "passe". D'autre part, la question sous-jacente est la suivante : quelle quantité d'informations un système d'inférence est-il capable d'apprendre ? Nous n'avons pas la réponse à cette question, mais nous avons posé les fondements d'un tel système.

Dans ce chapitre de perspectives, nous proposons simplement une approche différente du problème d'analyse de la structure. Cette approche repose sur l'utilisation non plus des modèles de Markov cachés mais des modèles de segments. Nous commençons par introduire le principe des modèles de segments. Puis nous en proposons une mise en œuvre pour la structuration d'une vidéo de sport, inspirée de nos précédents travaux.

5.1 Présentation des modèles de segments

5.1.1 Limitations des modèles de Markov cachés

La modélisation par modèles de Markov cachés présente quelques limitations. La première limitation, et probablement la faiblesse majeure des HMMs conventionnels, concerne la modélisation des durées. En effet, pour un HMM d'ordre 1, un modèle de durée $p_i(d)$ exponentiel est inhérent à chaque état. Il est défini pour chaque état de manière implicite par les probabilités de transitions (5.1). Soit a_{ii} la probabilité de bouclage sur l'état cou-

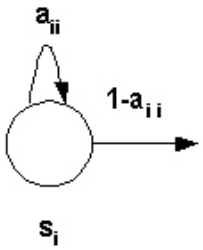
rant, et $(1-a_{ii})$ la probabilité de passage à un autre état, alors la probabilité que d trames exactement aient été émises par l'état i, puis que le processus quitte cet état, est évaluée comme la probabilité de la séquence d'observation

$$O = \underset{1}{\{S_i} \quad \underset{2}{S_i} \quad \underset{3}{S_i} \quad \ldots \quad \underset{d}{S_i} \quad \underset{d+1}{S_j \neq S_i\}}$$

soit :

$$p_i(d) = a_{ii}^{d-1}(1 - a_{ii}) \tag{5.1}$$

La probabilité $p_i(d)$ suit une loi en décroissance exponentielle (cf. Fig. 5.1).



Etat d'une machine

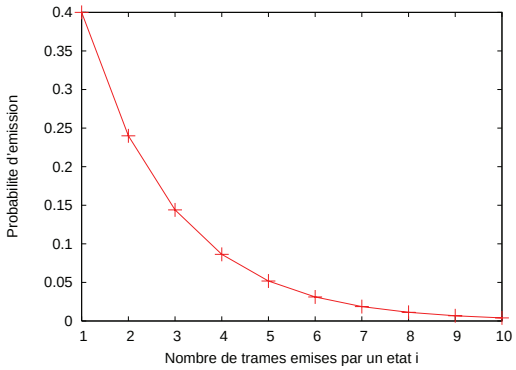


FIG. 5.1: Densité de probabilité de durée associée à un état d'une chaîne de Markov (pour $a_{ii} = 0,6$).

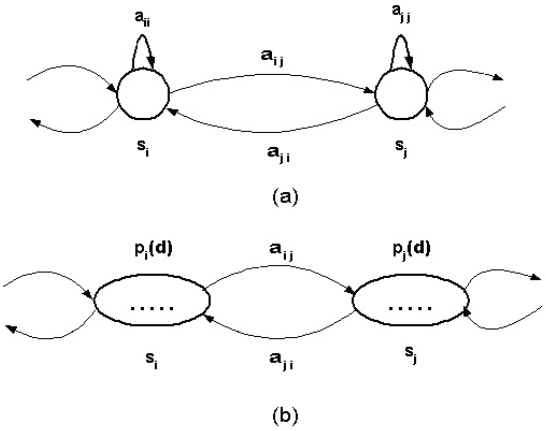


FIG. 5.2: Illustration des connexions inter-états dans un HMM (a) normal avec une densité exponentielle de durée d'un état, (b) de durée variable avec une densité de durée explicite.

Un modèle exponentiel n'est pas forcément approprié et, d'autre part, l'influence des probabilités de transitions dans le score étant souvent négligeable par rapport aux vraisemblances des observations, il n'a pas forcément un très grand rôle. Pour résoudre ce

problème de modélisation de la durée, on peut utiliser des HMMs d'ordre supérieur à 1 pour avoir, en théorie, un meilleur modèle de durée, ou ajouter explicitement une loi de durée aux états. Les deux solutions mentionnées n'ont cependant pas apporté d'amélioration significative des performances des HMMs. Une explication possible est, d'une part, que le modèle exponentiel n'est pas pertinent et, d'autre part, que les probabilités *a priori* n'ont pas une grande influence même lorsqu'on introduit une modélisation explicite de la durée.

5.1.2 Définition des modèles de segments

Pour remédier aux faiblesses de la modélisation par HMM, une famille de modèles, appelés modèles de segments ou modèles de trajectoires, a été introduite en 1996 par Ostendorf *et al.* [7]. Dans cette modélisation, les observations sont décrites comme des segments. Un modèle est associé à chaque segment, et la suite des segments est définie comme un processus markovien d'ordre 1. On a donc un segment plutôt qu'une seule trame associé à un état de la chaîne de Markov, comme l'illustre la figure 5.3. Un segment est modélisé par une loi de durée explicite ou implicite, $p(l|s)$, et par l'ensemble des densités d'émission $b(y_1, y_2, ..., y_l|l, s)$.

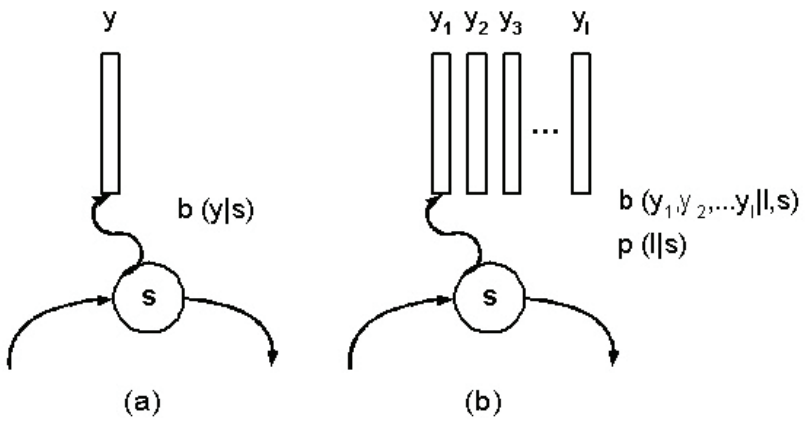


FIG. 5.3: Représentation de la modélisation par modèles de segments (d'après [7]). (a) principe des HMM, (b) principe des modèles de segments.

Une technique communément utilisée consiste à diviser un segment en régions et à associer à chaque région une densité d'émission dont les paramètres sont constants. Il est alors nécessaire de définir une correspondance entre les trames et les régions. La correspondance se fait soit de manière déterministe, par l'intermédiaire d'une table de correspondance ou d'une trajectoire échantillonnée, soit de manière dynamique en utilisant des algorithmes de programmation dynamique pour mettre en correspondance une trajectoire donnée avec un nombre de régions fixé à l'avance.

5.1.3 Discussion

Les modèles de trajectoires ont été utilisés avec succès dans des systèmes de reconnaissances de grand vocabulaire. La parole est décrite comme une suite de segments. Des segments de la taille d'un phone sont en général utilisés, bien qu'il soit possible de modéliser des segments de taille quelconque.

Les modèles de segments peuvent être vus comme une version des modèles de Markov cachés de plus grande dimension, dans le sens où les états génèrent des séquences aléatoires plutôt qu'un seul vecteur d'observation. Puisqu'il s'agit d'une généralisation des HMMs, les algorithmes standards d'apprentissage et de décodage peuvent être facilement étendus au cas des modèles de segments, avec cependant un coût de calcul beaucoup plus élevé étant donné l'augmentation de la taille de l'espace des états et du nombre de paramètres. Ces modèles pourraient être utilisés pour représenter la structure d'une vidéo.

5.2 Application à la structuration

Nous pouvons conserver la topologie des unités logiques et du modèle de Markov hiérarchique présentés aux chapitres 2 et 4. Ce sont les états émetteurs qui sont transformés en modèles de segments, et donc la représentation des observations et les lois de probabilités associées qui sont modifiées.

Les vecteurs d'observations ne sont plus des attributs du plan, mais des attributs de l'image pour la vidéo et des trames pour l'audio. Le schéma du système de structuration correspondant est représenté à la figure 5.4.

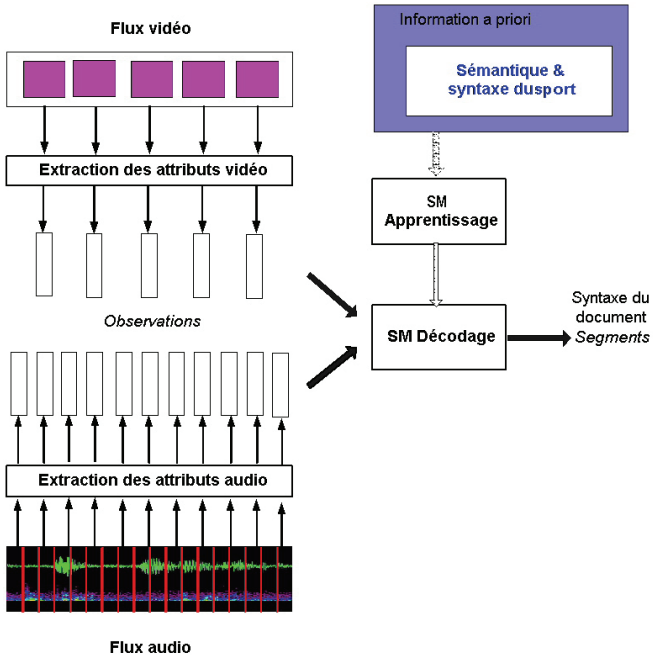


FIG. 5.4: Schéma du système de structuration par modèles de segments (SM).

Nous allons détailler les états du modèle de segments et les densités de probabilités d'observations associées.

5.2.1 Description d'un état et des observations associées

Un état du modèle de segment modélise un plan au sens de la vidéo. Cet état génère donc une séquence d'observation de longueur variable correspondant à un plan vidéo. Afin de prendre en compte la modalité audio, une séquence d'observation audio est également associée à l'état. Chaque état génère donc deux segments d'observations, l'un vidéo et l'autre audio (Fig. 5.5).

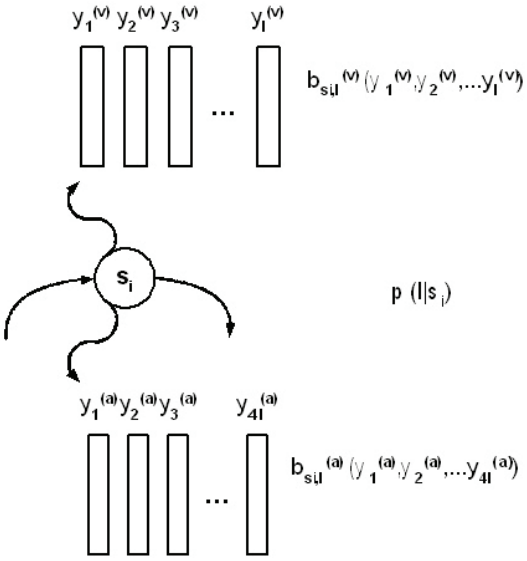


FIG. 5.5: Séquences d'observations vidéo et audio de longueur variable l générées par un état du modèle de segment.

La fréquence d'échantillonnage du flux vidéo est de 25 images par seconde, soit une image toutes les 40 ms. Concernant l'audio, on ne sait modéliser que des processus stationnaires. Il est donc nécessaire de modéliser des petites plages sur lesquelles on peut supposer que le processus est stationnaire. La fréquence d'échantillonnage est de 10 ms. Afin de synchroniser les deux flux, on considère que pour un vecteur d'observation vidéo, 4 vecteurs d'observations audio sont extraits (Fig. 5.6).

Soit une séquence d'observations $y_{t_1}^{t_2} = \{y_{t_1}, y_{t_1+1}, \ldots, y_k, \ldots, y_{t_2}\}$ de longueur $l_i = t_2 - t_1$ aléatoire, alors la probabilité de la séquence d'observations sachant l'état s_i est donnée par la probabilité jointe de la séquence vidéo $y_{t_1}^{t_2(v)}$ et de la séquence audio $y_{t_1}^{t_2(a)}$ conditionnellement à l'état s_i :

$$\begin{aligned} p(y_{t_1}^{t_2}, l_i|s_i) &= p(y_{t_1}^{t_2}|l_i, s_i) \cdot p(l_i|s_i) \\ &= p(y_{t_1}^{t_2(v)}, y_{4t_1}^{4t_2(a)}|l_i, s_i) \cdot p(l_i|s_i) \end{aligned} \tag{5.2}$$

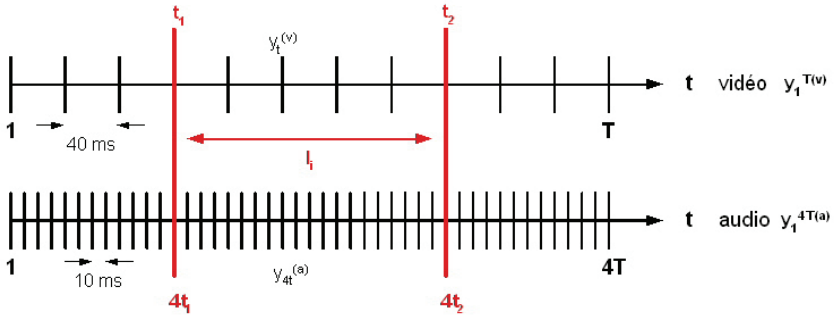


FIG. 5.6: Echantillonnage du flux audio et vidéo.

Les séquences vidéo et audio sont supposées indépendantes l'une de l'autre et indépendantes par rapport à la longueur l_i :

$$p(y_{t_1}^{t_2}, l_i|s_i) = p(y_{t_1}^{t_2(v)}|s_i) \cdot p(y_{4t_1}^{4t_2(a)}|s_i) \cdot p(l_i|s_i) \tag{5.3}$$

En revanche, les vecteurs $y_1^{(v)}, y_2^{(v)}, ..., y_T^{(v)}$ ne sont pas supposés indépendants les uns des autres, ce qui distingue ce modèle d'un modèle de Markov caché. La probabilité d'une séquence $y_{t_1}^{t_2}$ sachant l'état s_i s'exprime alors comme :

$$p(y_{t_1}^{t_2}, l_i|s_i) = \prod_{j=1}^{J} p(f_j(y_{t_1}^{t_2(v)})|s_i) \cdot \prod_{k=1}^{K} p(f_k(y_{4t_1}^{4t_2(a)})|s_i) \cdot p(l_i|s_i) \tag{5.4}$$

où J est le nombre de caractéristiques extraites pour chaque vecteur vidéo, K le nombre de caractéristiques extraites pour chaque vecteur audio et $f_{j/k}(y_{t_1}^{t_2})$ est une fonction des vecteurs de la séquence définie pour une caractéristique donnée.

5.2.2 Densités de probabilités d'observations

Pour chaque image, les attributs extraits sont un histogramme couleur, les couleurs dominantes et l'activité moyenne dans l'image. Pour chaque frame audio, un vecteur acoustique contenant l'énergie et les coefficients cepstraux est extrait.

5.2.2.1 Densités de probabilités associées aux observations vidéo

Couleur

Si l'on considère qu'un état du modèle de segment modélise un plan de la vidéo, les couleurs des vecteurs d'observations doivent être homogènes au sein d'un état. En s'inspirant des méthodes classiques de segmentation temporelle, $f_h(y_{t_1}^{t_2})$ est une fonction des histogrammes des vecteurs de la séquence définie par :

$$f_h : \{y_{t_1}, ..., y_k, ...y_{t_2}\} \mapsto \{h_{t_1+1} - h_{t_1}, ..., h_{k+1} - h_k, ...h_{t_2+1} - h_{t_2}\} \tag{5.5}$$

où h_k est l'histogramme associé au vecteur y_k. On note $f_h[k] = f_h(y_k) = h_{k+1} - h_k$. On définit alors une loi de probabilité binaire associée à f_h :

$$p(f_h(y_{t_1}^{t_2})|s_i) = \begin{cases} 1-\varepsilon & \text{si} f_h[k] < Th_{cut}, \forall k \in [t_1, t_2[\\ & \text{et} f_h[t_2] > Th_{cut} \\ \varepsilon & \text{sinon} \end{cases}$$

Couleurs dominantes

Dans la mesure où on cherche à caractériser les plans représentants une vue globale du terrain, on ne considère non plus seulement l'homogénéité des couleurs de la séquence générée par un état, mais la similarité des couleurs dominantes de cette séquence par rapport à une couleur dominante modèle du terrain $\bar{c}_{terrain}$. $f_d(y_{t_1}^{t_2})$ est une fonction des couleurs dominantes des vecteurs de la séquence définie par :

$$f_d : y_k \mapsto d(c_k, \bar{c}_{terrain}), \forall k \in [t_1, t_2] \tag{5.6}$$

où c_k est le vecteur de couleurs dominantes associé au vecteur y_k, et $d(c_k, \bar{c}_{terrain})$ est la mesure de similarité visuelle entre c_k et le modèle du terrain $\bar{c}_{terrain}$.

La probabilité d'observation associée à un état s_{autre} ne représentant pas une vue globale du terrain est alors :

$$p(f_d(y_k)|s_{autre}) = 1 - p(f_d(y_k)|s_{terrain}) \tag{5.7}$$

On suppose que les vecteurs $f_d(y_k)$ sont indépendants les uns des autres. Alors la probabilité d'une séquence $y_{t_1}^{t_2(a)}$ sachant l'état s est :

$$p(f_d(y_{t_1}^{t_2(v)})|s) = \prod_{k=t_1}^{t_2} p(f_d(y_k^{(v)})|s) \tag{5.8}$$

L'unique application de cette probabilité d'observation conduira à un découpage de la vidéo en scènes de jeu (représentées par des vues globales du terrain) et de non-jeu, sans tenir compte du découpage en plan de la vidéo. Concernant la structuration d'une vidéo de tennis, cette condition est suffisante. Mais pour élargir l'application à d'autres types de vidéos, les deux distributions précédemment explicitées doivent être prise en compte :

- la densité de probabilité associée à l'homogénéité des couleurs dans une séquence générée par un état, qui conduit à une segmentation temporelle de la vidéo ;
- la densité de probabilité associée à la similarité visuelle entre les observations générées par un état et un modèle visuel, qui permet de caractériser le contenu du plan.

Activité

Le mouvement de caméra est considéré à travers l'activité moyenne d'un segment :

$$f_a(y_{t_1}^{t_2}) = \frac{1}{t_2 - t_1} \sum_{k=t_1}^{t_2} \sqrt{u_k^2 + v_k^2} \tag{5.9}$$

où u_k et v_k sont respectivement les vecteurs de mouvement MPEG horizontaux et verticaux, associés à une observation y_k.

5.2.2.2 *Densités de probabilités associées aux observations audio*

Chaque état est caractérisé par un mélange de gaussiennes à M composantes $(w_i, \mu_i, \Sigma_i)_{1 \leq i \leq M}$, avec μ_i la moyenne, Σ_i la matrice de covariance et w_i le poids affecté à la gaussienne i, les w_i satisfaisant $\sum_{i=1}^{M} w_i = 1$. Pour chaque gaussienne, les paramètres w_i, μ_i et Σ_i sont estimés à partir de l'ensemble des vecteurs acoustiques des données d'apprentissage. La loi de densité d'un tel mélange de gaussiennes pour un vecteur acoustique $y_k^{(a)}$ de dimension d est définie par :

$$p(y_k^{(a)}) = \sum_{i=1}^{M} w_i \; N(y_k^{(a)}, \mu_i, \Sigma_i) \tag{5.10}$$

où $N(y_k^{(a)}, \mu_i, \Sigma_i)$ est la densité de la gaussienne i :

$$N(y_k^{(a)}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp(-\frac{1}{2}(y_k^{(a)} - \mu_i)^T \Sigma_i^{-1} (y_k^{(a)} - \mu_i)) \tag{5.11}$$

Etant donnée une séquence acoustique $y_{t_1}^{t_2(a)}$, les vecteurs $\{y_k\}_{k \in [t_1, t_2]}$, sont supposés indépendants les uns des autres. La probabilité d'une séquence $y_{t_1}^{t_2(a)}$ sachant l'état s est alors :

$$\begin{aligned} p(y_{t_1}^{t_2(a)}|s) &= \prod_{k=t_1}^{t_2} p(y_k^{(a)}|s) \\ &= \prod_{k=t_1}^{t_2} \sum_{i=1}^{M} w_i \; N(y_k^{(a)}, \mu_i, \Sigma_i) \end{aligned} \tag{5.12}$$

5.2.3 Apprentissage et décodage

En substituant les probabilités précédemment présentées dans l'équation 5.4, la probabilité d'une séquence $y_{t_1}^{t_2}$ sachant l'état s_i s'exprime finalement par :

$$p(y_{t_1}^{t_2}, l_i|s_i) = p(f_h(y_{t_1}^{t_2(v)})|s_i) \cdot p(f_d(y_{t_1}^{t_2(v)})|s_i) \cdot p(f_a(y_{t_1}^{t_2(v)})|s_i) \cdot \prod_{k=4t_1}^{4t_2} p(y_k^{(a)}|s_i) \cdot p(l_i|s_i) \tag{5.13}$$

L'algorithme de décodage utilisé pour les modèles de segments est similaire à celui utilisé pour les HMMs, à ceci près qu'un état inclut à la fois le label et la durée. En d'autre mots, un état d'un modèle de segment est $q = (s, l) \in \mathcal{S} \times \mathcal{L} = \mathcal{Q}$, tandis que l'état d'un HMM est représenté par $q = s \in \mathcal{S} = \mathcal{Q}$. Le décodage consiste à segmenter une séquence d'observations y_1^T en une séquence d'états q_1^N. Cette segmentation étant spécifiée de manière unique par la séquence des longueurs des segments $l_1^N = \{l_1, ..., l_N\}$, on a :

$$p(y_1^T|q_1^N) = \sum_{l_1^N} p(y_1^T, l_1^N|q_1^N) = \sum_{l_1^N} p(y_1^T|l_1^N, q_1^N) p(l_1^N|q_1^N) \tag{5.14}$$

avec :

$$p(y_1^T|l_1^N, q_1^N) = \prod_{i=1}^{N} p(y_{t_{(i-1)}}^{t_{(i)}}|l_i, q_i) \tag{5.15}$$

$$p(l_1^N|q_1^N) = \prod_{i=1}^{N} p(l_i|q_i, l_{i-1}, q_{i-1}) \tag{5.16}$$

$t(i)$ représente la fin du $i$ème segment et $l_i = t(i) - t(i-1)$ est la longueur du segment. La reconnaissance des segments consiste donc à trouver :

$$\hat{q}_1^N = \arg\max_{q_1^N} p(q_1^N|y_1^T) = \arg\max_{q_1^N} \frac{p(y_1^T|q_1^N)p(q_1^N)}{p(y_1^T)} \tag{5.17}$$

ce qui s'écrit de façon plus précise :

$$(\hat{N}, \hat{q}_1^{\hat{N}}) = \arg\max_{N, q_1^N} (\max_{l_1^N} \; p(y_1^T|l_1^N, q_1^N)\; p(l_1^N|q_1^N)\; p(q_1^N)) \tag{5.18}$$

La différence clé entre les algorithmes de décodage des modèles de segments et des HMMs est l'évaluation explicite de différentes segmentations, qui ajoute une dimension supplémentaire à la recherche par programmation dynamique. Pour une séquence de longueur T, une recherche exhaustive considérera environ 2^T segmentations. Le nombre de segments à évaluer est diminué en réduisant l'espace de recherche. Une stratégie possible est d'introduire des contraintes sur la taille des segments recherchés en définissant $\rho(t, j)$, l'ensemble des longueurs de segments possibles pour un état s_j se terminant à t.

L'algorithme de programmation dynamique pour les modèles de segments s'écrit alors :

1. Initialisation : $t = 1$

$$\begin{array}{lcll} \delta_1(i) & = & \log p(y_1|l, i)p(l|i)p(i) & \forall i \in \mathcal{S}, l = 1 \\ \psi_1(i) & = & 0 & \end{array} \tag{5.19}$$

2. Récursion : $t = 2, ..., T$

$$\begin{array}{lcll} \delta_t(i) & = & \max_{j \in \mathcal{S}, \tau \in \rho(t,i)} \delta_\tau(i) + \log[p(y_{\tau+1}^t|l_\tau, i)p(l_\tau|i)p(i|j)] & \forall i \in \mathcal{S}, l_\tau = t - \tau \\ \psi_t(i) & = & \arg\max_{j \in \mathcal{S}, \tau \in \rho(t,i)} [\delta_\tau(j)a_{ij}] + \log[p(y_{\tau+1}^t|l_\tau, i)p(l_\tau|i)p(i|j)] & \forall i \in \mathcal{S}, l_\tau = t - \tau \end{array} \tag{5.20}$$

3. Terminaison :

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{5.21}$$

$$\hat{q} = \arg\max_{i \in \mathcal{S}} [\delta_T(i)] \tag{5.22}$$

4. Reconstruction :

$$(\hat{q}_t', t') = \psi_t(\hat{q}_t) \tag{5.23}$$

5.3 Conclusion

Dans ce chapitre, nous avons proposé une solution alternative aux modèles de Markov cachés sous la forme des modèles de segments. Cette méthode est plus complexe à mettre en œuvre et les temps de calcul du décodage sont plus importants, mais elle permet de réaliser simultanément la segmentation et la structuration de la vidéo sans découpage, ni traitements préalables. Les modèles de segments offrent également un cadre commun à l'intégration d'indices sonores tout en s'affranchissant des problèmes de synchronisation.

Cette méthode demande bien sûr à être validée expérimentalement et constitue l'une des évolutions et des perspectives possibles de notre travail. Nous devons encore souligner ici que nous n'avons pas introduit l'information, pourtant capitale pour la structuration hiérarchique, de la position du joueur. Nous considérions jusqu'à présent la position du joueur à l'engagement d'une balle. Dans une approche image, il faut prendre en compte l'évolution de la position du joueur au cours du plan. Il est nécessaire d'envisager pour l'application aux modèles de segment une solution à la représentation de cette information.

Conclusion générale

Nous avons abordé le problème de la reconnaissance de la structure d'une vidéo dont le contenu est fortement structuré. Nous avons présenté un modèle statistique audio-visuel pour l'analyse de la structure temporelle d'une vidéo de tennis. Le cadre général de la modélisation est celui des modèles de Markov cachés.

Dans cette conclusion générale, nous présentons tout d'abord une synthèse des travaux effectués puis nous esquissons plusieurs perspectives dans le prolongement de cette étude.

Synthèse des travaux effectués

Les principales contributions de nos travaux concernent l'exploitation de l'information *a priori* pour extraire une connaissance de haut-niveau du contenu d'une vidéo donnée. Plus précisément, nous avons abordé les aspects suivants :

- la **modélisation statistique de la structure** d'un document par des modèles de Markov cachés qui intègrent l'information *a priori* sur le contenu de la vidéo et les règles d'édition ;
- l'**intégration d'informations multimodales** pour décrire le contenu de la vidéo ;
- la **recherche d'une structure hiérarchique** de haut-niveau proposant un découpage sémantique de la vidéo.

Modélisation statistique de la structure

Nous avons présenté un modèle statistique pour la macro-segmentation d'une vidéo de tennis en unités logiques. Nous avons défini ces unités logiques de façon à ce que des informations sémantiques et structurelles puissent être déduites de la macro-segmentation. Le modèle proposé repose sur l'analyse de l'entrelacement temporel des plans et sur la caractérisation du contenu des plans en type de vues, à partir d'informations visuelles. Les attributs utilisés sont la longueur des plans, le type de transitions entre les plans et une similarité visuelle définie entre l'image-clé du plan et une image-clé de référence extraite automatiquement de la vidéo. Les unités logiques sont modélisées par des modèles de Markov cachés qui intègrent les informations *a priori* sur le contenu de la vidéo et les règles de production. Un algorithme de programmation dynamique réalise simultanément la classification et la segmentation d'une séquence de tennis en unités logiques. Nous avons validé ce modèle sur un ensemble de vidéos de tennis provenant de différents tournois et de différents pays.

Intégration d'informations audio-visuelles

Nous avons proposé une méthode d'intégration d'informations sonores. Les modèles de Markov cachés fournissent un cadre probabiliste efficace pour l'intégration de données multimodales. Les informations audio utilisées sont de moyen niveau. Elles sont représentées en terme d'absence ou de présence d'une classe audio prédéterminée dans le plan vidéo. Il s'agit d'une représentation simple mais efficace. En effet, les expérimentations montrent que l'intégration d'informations multimodales augmente les performances de la classification lorsque les caractéristiques utilisées sont de bonne qualité. Nous avons constaté l'importance de la qualité du processus de segmentation de la bande sonore en classes, dans ce cadre de représentation. Les erreurs en provenance de l'extraction des caractéristiques se répercutent au niveau de la structuration.

Recherche de la structure hiérarchique

La segmentation en unités logiques ne permet pas d'accéder directement à la structure hiérarchique des sports étudiés. Nous avons proposé de modéliser et de segmenter l'ensemble de cette structure hiérarchique par des modèles de Markov cachés hiérarchiques. L'ensemble des règles et des informations *a priori* que le système doit inférer devient alors encore plus complexe. Les informations audio-visuelles peu spécifiques que nous utilisions jusqu'alors se révèlent insuffisantes. Nous intégrons un indice de haut-niveau spécifique au tennis, qui permet à l'algorithme de programmation dynamique de retrouver son chemin dans le graphe d'états. Nous obtenons alors une structuration dense de l'ensemble de la vidéo de très-haut niveau.

Perspectives

Les perspectives que nous envisageons dans le prolongement de ces travaux de thèse s'articulent autour de deux axes de recherche. En premier lieu, il nous semble intéressant d'approfondir la modélisation par modèles de Markov cachés proposée. D'autre part, une approche différente du problème d'analyse de la structure se dégage de la modélisation précédente. Cette approche repose sur l'utilisation des modèles de segments.

Modélisation par modèles de Markov cachés

Une première perspective est l'extraction d'attributs audio-visuels plus complexes en entrée du modèle. La description sonore en terme d'absence ou de présence d'une classe audio dans le plan est un modèle de représentation très simple qui, s'il est efficace, peut être amélioré. Par exemple, il n'y a pas de mesure qualifiant l'importance de la présence de classe dans les plans. L'évolution temporelle du signal audio dans un plan peut également être prise en compte. Des informations visuelles telles que l'identification des scores incrustés dans les vidéos peuvent être intégrées au système, soit comme une alternative à l'extraction des joueurs et du terrain, soit afin de compléter le processus de structuration. La mauvaise qualité de l'image rend la reconnaissance du score incrusté difficile. Cependant, cette reconnaissance peut être guidée par l'intégration au modèle de l'évolution connue du score au cours du match. Il faut alors prendre en compte que l'apparition du score dans l'image est postérieur à l'événement.

Une autre piste à explorer est l'intégration de la connaissance *a posteriori* du score du match. Connaissant le résultat du match, comment intégrer cette information pour guider le processus de structuration ?

Enfin, la structuration de la vidéo peut être utilisée comme filtrage pour la détection d'événements. Les segments analysés seront sélectionnés parmi les phases de jeu pertinentes mises en avant par la structuration. Le recalage du terrain et le suivi des joueurs permettent d'identifier ultérieurement les phases de jeu (service, volée...). Ces même attributs peuvent être utilisés en retour pour corriger la segmentation de la structure. On peut alors envisager un système d'indexation complet, depuis l'analyse de la structure jusqu'à la détection d'événements particuliers.

Modélisation par modèles de segments

Les modèles de segments sont une alternative intéressante aux modèles de Markov cachés. Ils modélisent l'évolution temporelle des observations pour un état donné. De plus cette méthode permettrait de réaliser simultanément la segmentation et la structuration de la vidéo à partir des attributs audio-visuels bas-niveau sans segmentation, ni traitements préalables. Les modèles de segments offrent également un cadre commun à l'intégration multimodale tout en s'affranchissant des problèmes de synchronisation.

Annexes

Annexe A

Rappel des règles du tennis

Le tennis est un jeu qui oppose deux joueurs et qui consiste à faire passer une balle une fois de plus que l'adversaire au-dessus d'un filet, à l'aide d'une raquette, sans la faire aller plus loin que les lignes qui délimitent le terrain.

A.1 Le terrain

Un terrain de tennis est délimité par des lignes blanches. La surface de terrain où la balle peut rebondir sans qu'elle soit considérée comme faute est celle comprise entre les lignes de fond de court, lignes extérieures parallèles au filet, et les lignes latérales, perpendiculaires au filet (les lignes les plus à l'extérieur pour un match en double et celles qui sont à l'intérieur pour un match en simple : les parties du terrain situées entre ces deux types de lignes latérales sont les couloirs). On trouve aussi les lignes de services qui sont parallèles aux lignes de fond de court et coupent chaque moitié de terrain en deux parties égales. Enfin, parallèles aux lignes latérales, il existe des lignes au centre du terrain qui servent à la délimitation des carrés de service (Fig. A.1). Il existe différentes surfaces de terrain : la terre battue, les bétons poreux, les résines imperméables et le gazon, qui ont chacune leur propre couleur.

A.2 Le jeu

Le tennis repose sur l'échange de la balle entre les adversaires. La balle doit être renvoyée après un rebond ou avant qu'elle ait rebondit (on parle alors de reprise de volée) et doit retomber dans le côté opposé dans les limites du terrain. On marque le point lorsque l'adversaire ne réussit pas à renvoyer la balle, ou qu'il la renvoie hors des limites du terrain, ou qu'elle retombe dans son propre camp.

Un *échange* est engagé par un *service*. Le service se fait à partir de l'arrière de la ligne de fond de court. La balle d'engagement doit rebondir à l'intérieur d'un carré de service, sinon elle est considérée comme faute. Le joueur sert alternativement sur l'un et l'autre des carrés de services. Il a droit à deux services : on parle de *premier service* et de deuxième service. S'il fait deux fautes successives, il y a double faute et il perd le point. Lorsqu'une balle de service touche le filet en retombant dans le carré de service, on dit qu'elle est *let* et

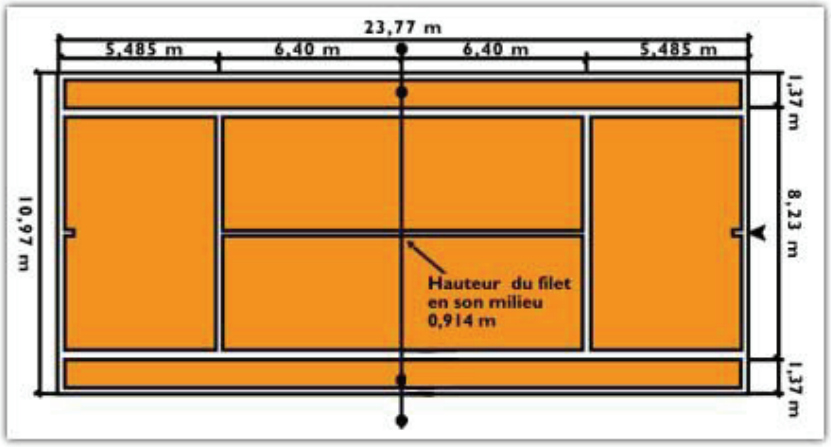


FIG. A.1: Terrain de tennis : au milieu se situe le filet, les lignes les plus à droite et à gauche sont les lignes de fond de court, les quatre carrés au centre de part et d'autre du filet sont les carrés de service, le bandes horizontales en haut et en bas sont les couloirs.

le service est rejoué. Lorsqu'une balle de service est bonne, mais que l'adversaire n'arrive pas à la renvoyer, on dit que le serveur a fait un *ace*.

A.3 Le déroulement du jeu

Le jeu doit être continu depuis le premier service jusqu'à ce que l'un des adversaires gagne le match. Un match de tennis se divise en *sets* (ou manches) qui eux-mêmes se subdivisent en *jeux* (Figure A.2). Les jeux sont servis à tour de rôle par les adversaires. Un jeu se joue normalement en 4 *points* gagnants : 15 pour le premier point remporté, puis 30, 40 et jeu pour les suivants. Lorsque les deux joueurs sont à 40 partout, le point suivant ne permet pas le gain du jeu. Celui qui remporte le point prend l'avantage et peut s'il remporte le point suivant gagner le jeu. Si c'est son adversaire qui remporte le point suivant, on dit que les deux joueurs sont à égalité ce qui ramène à la même situation qu'à 40 partout. Les sets se jouent généralement en 6 jeux gagnants avec deux jeux minimum d'écart avec l'adversaire. On gagne par exemple le set sur une marque de 6-4, 6-1 ou 7-5.

Les matchs de tennis se jouent généralement en 2 sets gagnants (3 sets gagnants pour les hommes pour certains tournois).

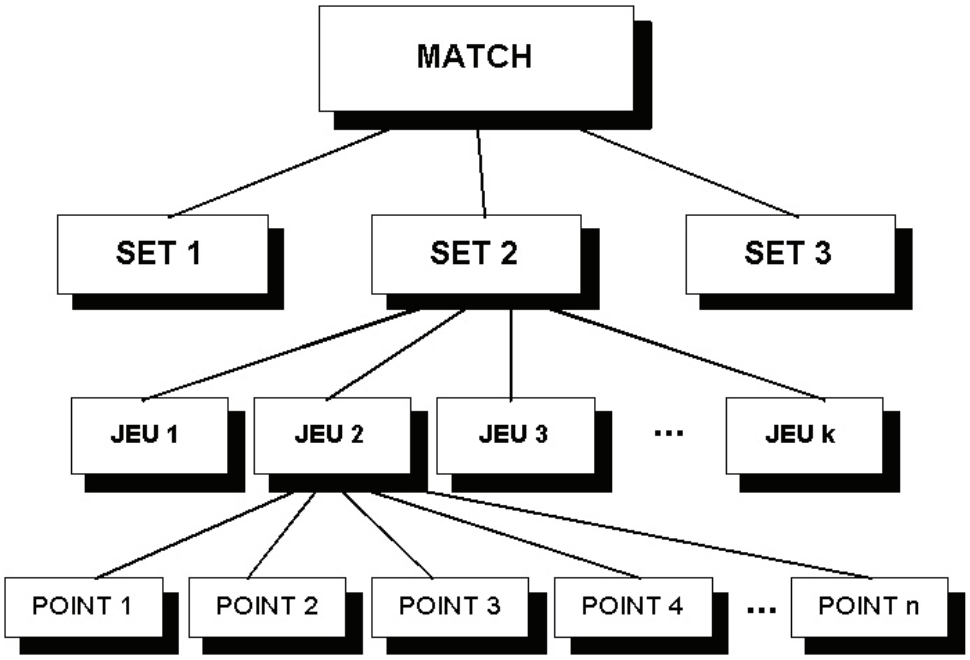


FIG. A.2: Structure intrinsèque d'un match de tennis.

Annexe B

Rappel des règles du baseball

Le baseball est un jeu qui oppose deux équipes de neuf joueurs qui deviennent alternativement équipe défensive et équipe attaquante. Les attaquants essayent de frapper avec une batte une balle lancée par le lanceur de l'équipe adverse. Les défenseurs essayent de rattraper la balle avant que les attaquants n'aient pu faire le tour du terrain en courant.

B.1 Le terrain

Le baseball se joue sur un terrain en forme de quart de cercle, divisé en champs : le champ intérieur ou diamant(Infield), et le grand champ ou champ extérieur (Outfield). Le champ intérieur a une forme carrée dont chaque sommet est marqué par une base. Celles-ci sont numérotées respectivement de 1 à 3, la dernière étant dénommée marbre ou maison (Home plate). Au centre du diamant se trouve le monticule sur lequel est placé une plaque où le *lanceur* prendra place (Fig. B.1).

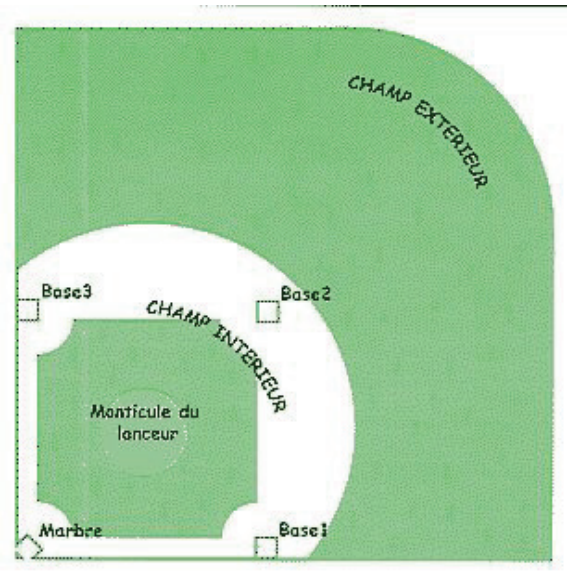


FIG. B.1: Terrain de baseball. En vert : gazon ; en blanc : stabilisé.

L'équipe défensive se positionne de la manière suivante :
- dans le champ intérieur : un joueur en 1ère base, un joueur entre la 1ère et la 2ème base, un joueur entre la 2ème et la 3ème base, un joueur en 3ème base, le lanceur sur le monticule au centre et le receveur derrière le marbre.
- dans le grand champ : trois joueurs se répartissent sur la surface du terrain.

L'équipe attaquante envoie ses joueurs à la frappe selon un ordre établi.

B.2 Le jeu

Le batteur se place sur le marbre. Le lanceur, du haut de son monticule, doit lancer les balles dans une zone bien déterminée appelée *zone de strike* (Figure B.2). Une balle qui passe dans cette zone est une bonne balle (*strike*). Un batteur a droit a trois essais, après quoi il est éliminé. Si le lanceur réalise quatre mauvais lancers (en dehors de la zone de strike), le batteur va automatiquement en 1ère base.

Le batteur devient coureur une fois qu'il a frappé la balle. Son objectif est alors de courir de base en base, dans l'ordre de leur numérotation, jusqu'à revenir au marbre, et cela avant que la défense rattrape la balle et l'achemine jusqu'à l'une des bases. Le tour complet des bases se fait généralement en plusieurs étapes. Un joueur devenu coureur ne peut pas se faire éliminer quand il est en contact avec une base. Mais si la balle arrive avant lui sur la base ou qu'il se fait toucher par un défenseur qui détient la balle entre deux bases, il est éliminé. Si l'attaquant réussi à rejoindre le marbre sans se faire éliminer, il marque un point.

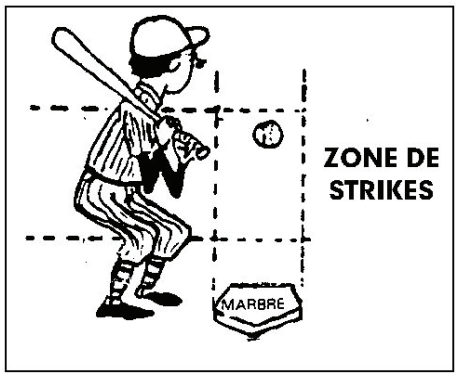


FIG. B.2: Zone de strikes.

B.3 Le déroulement du jeu

Lorsque les défenseurs ont éliminé 3 attaquants, il y a changement et ils passent en attaque. Le passage des deux équipes à la batte constitue une reprise ou manche (inning). Chaque manche est donc divisée en deux demie-manches où chaque équipe passe une fois à l'attaque et une fois en défense. Un match comprend 9 manches (Figure B.3).

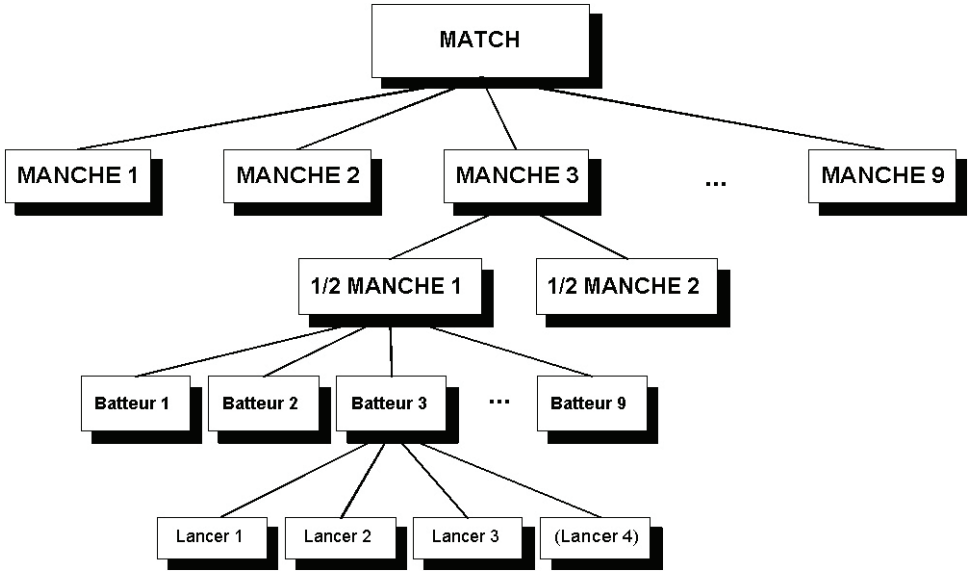


FIG. B.3: Structure intrinsèque d'un match de baseball.

Bibliographie

[1] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati. Automatic extraction and annotation of soccer video highlights. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[2] B. Li, J. Errico, H. Pan, and M.I. Sezan. Bridging the semantic gap in sports. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-5021, pages 314–326, San Jose, CA, January 2003.

[3] A. Ekin and A.M. Tekalp. Robust dominant color region detection and color-based applications for sports video. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[4] N. Nitta, N. Babaguchi, and T. Kitahashi. Story based representation for broadcasted sports video and automatic story segmentation. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[5] M. Xu, L-Y. Duan, C-S. Xu, and Q. Tian. A fusion scheme of visual and auditory modalities for event detection in sports video. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, Hong Kong, April 2003.

[6] C.G.M. Snoek and M. Worring. Time interval maximum entropy based event indexing in soccer video. In *IEEE International Conference on Multimedia and Expo (ICME03)*, volume 3, pages 481–484, Baltimore, MD, 2003.

[7] M. Ostendorf, V. Digalakis, and O. Kimball. From hmms to segment models : a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 4 :360–378, 1996.

[8] M. Xu, N.C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *IEEE International Conference on Multimedia and Expo (ICME03)*, volume 2, pages 281–284, July 2003.

[9] N. Dimitrova, H-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *IEEE Multimedia*, 9(3) :42–55, July-September 2002.

[10] R. Lienhart. Reliable transition detection in videos : A survey and practitioner's guide. *International Journal of Image and Graphics*, 1(3) :469–486, 2001.

[11] S. Lefevre. *Détection d'Evénements dans une Séquence Vidéo*. PhD thesis, Université de Tours, Décembre 2002.

[12] Y. Rui, T.S. Huang, and S. Mehrota. Constructing table-of-content for video. *Multimedia Systems*, 7(5) :359–368, September 1999.

[13] I. Koprinska and S. Carrato. Temporal video segmentation : a survey. *Signal Processing : Image Communication*, 16(5) :477–500, January 2001.

[14] M. Brand. The Inverse Hollywood Problem : From video to scripts and storyboards via causal analysis. In *AAAI/IAAI*, pages 132–137, 1997.

[15] P. Aigrain and P. Joly. The automatic real-time analysis of film editing and transition effects and its applications. *Computers & Graphics*, 18(1) :93–103, 1994.

[16] D. Swanberg, C-F. Shu, and R. Jain. Knowledge guided parsing in video databases. In *Electronic Imaging : Science and Technology*, volume SPIE-1908, pages 13–24, San Jose, California, February 1993.

[17] C.G.M. Snoek and M. Worring. Multimodal video indexing : A review of the state-of-the-art. *Multimedia Tools and Applications*, 2003. to appear.

[18] H. Zhang, S-Y. Tan, S.W. Smoliar, and G. Yihong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(6) :256–266, 1995.

[19] S. S. Intille and A. F. Bobick. Closed-world tracking. In *Proc. of the Fifth International Conference on Computer Vision (ICCV95)*, pages 672–678, June 1995.

[20] Y. Gong, L.T. Sin, C.H. Chuan, H.J. Zhang, and M. Sakauchi. Automatic parsing of TV soccer programs. In *International Conference on Multimedia Computing and Systems*, pages 167–174, May 1995.

[21] D. Yow, B. Yeo, M. Yeung, and G. Liu. Analysis and presentation of soccer highlights from digital video. In *Proc. of 2nd Asian Conference on Computer Vision*, pages 499–503, Singapore, December 1995.

[22] Y. Onho, J. Miura, and Y. Shirai. Tracking players and estimation of the 3d position of a ball in soccer games. In *IEEE International Conference on Pattern Recognition (ICPR00)*, pages 145–148, 2000.

[23] S. Choi, Y. Seo, H. Kim, and K-S. Hong. Where are the ball and players ? : Soccer game analysis with color-based tracking and image mosaick. Technical report, Dept. of EE, Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang, 790-784, Republic of Korea, 1997.

[24] T. Kawashima, K. Tateyama, T. Iijima, and T. Aoki. Indexing of baseball telecast for content-based video retrieval. In *IEEE International Conference on Image Processing (ICIP98)*, volume 1, pages 871–875, October 1998.

[25] D. Zhong and S-F. Chang. Structure analysis of sports video using domain models. In *IEEE International Conference on Multimedia and Expo (ICME01)*, Tokyo,Japan, August 2001.

[26] G. Sudhir, J. C. M. Lee, and A. K. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proc. Of IEEE Workshop on Content-Based Access of Image and Video Databases*, Bombay, January 1998.

[27] L-Y. Duan, M. Xu, and Q. Tian. Semantic shot classification in sports video. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-5021, pages 300–313, January 2003.

[28] B. Li and M.I. Sezan. Event detection and summarization in american football broadcast video. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-4676, pages 202–213, 2002.

[29] H. Denman, N. Rea, and A. Kokaram. Content-based analysis for video from snooker broadcasts. In *Proc. Int'l Conf. on Image and Video Retrieval*, volume 2383, pages 198–205, London, UK, July 2002. Springer, Lecture Notes in Computer Science.

[30] R. Dayhot, A. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, Hong Kong, April 2003.

[31] P. Xu, L. Xie, S-F. Chang, A. Divakaram, A. Vetro, and H. Sun. Algorithms and system for segmentation and structure analysis in soccer video. In *IEEE International Conference on Multimedia and Expo (ICME01)*, pages 928–931, August 2001.

[32] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7) :796–807, July 2003.

[33] A. Ekin and A.M. Tekalp. A framework for tracking and analysis of soccer video. In *IS&T/SPIE Visual Communications and Image Processing (VCIP02)*, San Jose, CA, January 2002.

[34] H-S. Yoon, Y-I. J. Bae, and Y-K. Yang. A soccer image sequence mosaicking and analysis method using line and advertisement board detection. *ETRI Journal*, 24(6) :443–454, December 2002.

[35] N. Vandenbroucke. *Segmentation d'Images Couleur par Classification de Pixels dans des Espaces d'Attributs Colorimétriques Adapté. Application à l'Analyse d'Images de Football.* PhD thesis, Université de Lille 1, Décembre 2000.

[36] H. Miyamori and S-I. Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *IEEE Fourth International Conference on Automatic Face and Gesture Recognition*, pages 320–325, Grenoble, France, March 2000.

[37] G.S. Pingali, Y. Jean, and I. Carlbom. Real-time tracking for enhanced tennis broadcasts. In *Proc. Of IEEE Computer Vision and Pattern Recognition (CVPR98)*, pages 260–265, 1998.

[38] A. Gueziec. Tracking pitches for broadcast television. In *IEEE Computer*, volume 35, pages 38–43, March 2002.

[39] H. Miyamori. Automatic annotation of tennis action for content-based retrieval by integrated audio and visual information. In *(CIVR03)*, pages 331–341, 2003.

[40] W. Zhou, A. Vellaikal, and C.-C. J. Kuo. Rule-based video classification system for basketball video indexing. In *Proc. ACM International Multimedia Conference*, pages 213–216, Los Angeles, California, November 2000.

[41] D. Zhong. *Segmentation, Index and Summarization of Digital Video Content.* PhD thesis, Columbia University, 2001.

[42] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2) :52–60, 2002.

[43] C. Wu, Y-F. Ma, H-J. Zhang, and Y-Z. Zhong. Events recognition by semantic inference for sports video. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[44] P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with Hidden Markov Models. In *Proc. of IEEE International Conference on Image Processing (ICIP02)*, Rochester, NY, USA, September 2002.

[45] H. Lu and Y-P. Tan. Sports video analysis and structuring. In *Proc. Of IEEE Workshop on Multimedia Signal Processing*, Cannes,France, October 2001.

[46] M.H. Lee, S. Nepal, and U. Srinivasan. Edge-based semantic classification of sports video sequences. In *IEEE International Conference on Multimedia and Expo (ICME03)*, July 2003.

[47] Y-P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Trans. on Circuits and Systems for Video Technology*, 10(1) :133–146, February 2000.

[48] K. Yoon, D. DeMenthon, and D. Doermann. Event detection from MPEG video in the compressed domain. In *IEEE International Conference on Pattern Recognition (ICPR00)*, Barcelona, Spain, 2000.

[49] L. Xie, S-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with Hidden Markov Models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, Orlando, FL, USA, May 2002.

[50] M. Lazarescu, S. Ventakesh, and G. West. On the automatic indexing of cricket using camera motion parameters. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[51] A. Kokaram and P. Delacourt. A new global motion estimation algorithm and its application to retrieval in sports events. In *Proc. Of IEEE Workshop on Multimedia Signal Processing*, Cannes,France, October 2001.

[52] N. Babaguchi, Y. Kawai, Y. Yasugi, and T. Kitahashi. Linking live and replay scenes in broadcasted sports video. In *Proc. Of ACM Multimedia Workshop on Multimedia Information Retrieval (MIR00)*, Los Angeles, CA, November 2000.

[53] V. Kobla, D. DeMenthon, and D. Doermann. Detection of slow-motion replays sequences for identifying sports video. In *Proc. Of IEEE Third Workshop on Multimedia Signal Processing*, pages 135–140, September 1999.

[54] H. Pan, P. Van Beek, and M.I. Sezan. Detection of slow-motion replays segments in sports video for highlights generation. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP01)*, pages 1649–1652, Salt Lake City, UT, May 2001.

[55] V. Kobla, D. DeMenthon, and D. Doermann. Identifying sports video using replay, text, and camera motion features. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-3972, pages 332–343, January 2000.

[56] H. Pan, B. Li, and M.I. Sezan. Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, pages 3385–3388, Orlando, FL, May 2002.

[57] D. Zhang and S-F. Chang. Event detection in baseball video using superimposed caption recognition. In *Proc. ACM International Multimedia Conference*, pages 315–318, Juan-les-pins, France, December 2002.

[58] D. Zhang, R.K. Rajendran, and S-F. Chang. General and domain-specific techniques for detecting and recognizing superimposed text in video. In *IEEE International Conference on Image Processing (ICIP02)*, Rochester, NY, September 2002.

[59] J.Kittler, K. Messer, W.J. Christmas, B. Levienaise-Obadia, and D. Koubaroulis. Generation of semantic cues for sports ideo annotation. In *IEEE International Conference on Image Processing (ICIP01)*, 2001.

[60] X. Gibert, H. Li, and D. Doermann. Sports video classification using hmms. In *IEEE International Conference on Multimedia and Expo (ICME03)*, Baltimore, MD, 2003.

[61] T. Lin and H-J. Zhang. Automatic video scene extraction by shot grouping. In *IEEE International Conference on Pattern Recognition (ICPR00)*, Barcelona, Spain, 2000.

[62] S'I. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga. Sports video categorizing method using camera motion parameters. In *IEEE International Conference on Multimedia and Expo (ICME03)*, Baltimore, MD, 2003.

[63] A. Ekin and A.M. Tekalp. Generic play-break event detection for summarization and hierarchical sports video analysis. In *IEEE International Conference on Multimedia and Expo (ICME03*, July 2003.

[64] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. on Multimedia*, 4(1) :68–75, March 2002.

[65] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using hmms. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[66] B. Li and M.I. Sezan. Semantic sports video analysis : Approaches and new applications. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[67] B. Li and M.I. Sezan. Event detection and summarization in sports video. In *IEEE Workshop on Content Based Access of Image and Video Databases*, December 2001.

[68] Q. Huang, Z. Liu, and A. Rosenberg. Automated semantic structure reconstruction and representation generation for broadcast news. In *IS&T/SPIE Storage and Retrieval for Image and Video Databases VII*, volume SPIE-3656, pages 50–62, San Jose, CA, January 1999.

[69] B. Merialdo, K. T. Lee, D. Luparello, and J. Roudaire. Automatic construction of personalized TV news programs. In *ACM Multimedia*, pages 323–331, San Jose, California, November 1999.

[70] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content-based indexing of images and video using face detection and recognition methods. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP01)*, Salt Lake City , Utah, May 2001.

[71] S'I. Satoh, T. Sato, M.A. Smith, Y. Nakamura, and T. Kanade. Name-it : Naming and detecting faces in news video. *IEEE Multimedia Magazine*, 6 :22–35, 1999.

[72] M. Bertini, A. Del Bimbo, and P. Pala. Content-based indexing and retrieval of TV news. *Pattern Recognition Letters*, 22(5) :503–516, 2001.

[73] W. Wolf. Hidden Markov Model parsing of video programs. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP97)*, volume 4, pages 2609–2611, April 1997.

[74] L. Chaisorn, T-S. Chua, and C-H. Lee. The segmentation of news video into story units. In *IEEE International Conference on Multimedia and Expo (ICME02)*, volume 1, pages 73–76, 2002.

[75] C-H. Demarty. *Segmentation et Structuration d'un Document Vidéo pour la Caractérisation et l'Indexation de son Contenu Sémantique : Application aux Journaux Télévisés.* PhD thesis, Ecole Nationale Supérieur des Mines de Paris, January 2000.

[76] B. Günsel, A.M. Ferman, and A.M. Tekalp. Temporal segmentation using unsupervised clustering and semantic object tracking. *Journal of Electronic Imaging*, 7(3) :592–604, July 1998.

[77] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of TV news for automatic topic retrieval. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP01)*, Salt Lake City , Utah, May 2001.

[78] A.G. Hauptmann and M.J. Witbrock. Story segmentation and detection of commercials in broadcast news. In *Advances in Digital Librairies Conference*, 1998.

[79] Y. Nakamura and T. Kanade. Spotting by association in news video. In *Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio*, March 1997.

[80] H. Sundaram. *Segmentation, Structure Detection and Summarization of Multimedia Sequences.* PhD thesis, Columbia University, 2002.

[81] A.M. Ferman and A.M. Tekalp. Probabilistic analysis and extraction of video content. In *IEEE International Conference on Image Processing (ICIP99)*, volume 2, pages 91–95, 1999.

[82] O. Javed, S. Khan, Z. Rasheed, and M. Shah. Visual content segmentation of talk & game shows. *International Journal of Computers and Applications*, June 2002.

[83] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[84] http ://www.fft.fr/competitions/regles.html.

[85] H.J Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1) :10–28, 1993.

[86] P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection.* John Wiley and Sons, New York, 1987.

[87] R. Lienhart, C. Kuhmunch, and W. Effelsberg. On the detection and recognition of television commercials. In *International Conference on Multimedia Computing and Systems*, pages 509–516, 1997.

[88] T. McGee and N. Dimitrova. Parsing TV program structures for identification and removal of non-story segments. In *IS&T/SPIE Storage and Retrieval for Image and Video Databases*, San Jose, CA, January 1999.

[89] E. Kijak, L. Oisel, and P. Gros. Temporal structure analysis of broadcast tennis video using Hidden Markov Models. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-5021, pages 289–299, San Jose, CA, January 2003.

[90] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for TV baseball programs. In *Proc. ACM International Multimedia Conference*, pages 105–115, Los Angeles, California, November 2000.

[91] D. Zhang and D. Ellis. Detecting sound events in basketball video archive. Technical report, Columbia University, New York City, NY 10025, 2001.

[92] H. Harb and L. Chen. Highlights detection in sports videos based on audio analysis. In *Third International Workshop on Content-Based Multimedia Indexing (CBMI03)*, pages 223–229, September 2003.

[93] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. *Traitement de la parole.* Presses Polytechniques et Universitaires Romandes, 2000. ISBN :2-88074-388-5.

[94] M. Carré and P. Philippe. Indexation audio : un état de l'art. *Annales des Télécommunications*, 55(9-10) :507–525, 2000.

[95] H. Wang, A. Divakaram, A. Veto, S-F. Chang, and H. Sun. Survey of compressed-domain features ussed in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, 2003. submitted.

[96] S. Pfeiffer and T. Vincent. Survey of compressed domain audio features and their expressiveness. In *IS&T/SPIE Storage and Retrieval for Media Databases*, volume SPIE-5021, pages 133–147, January 2003.

[97] S. Nepal. Automatic detection of 'goal' segments in basketball videos. In *Proc. ACM International Multimedia Conference*, pages 261–269, Ottawa, Ontario, Canada, October 2001.

[98] K. Kim, J. Choi, N. Kim, and P. Kim. Extracting semantic information from basketball video based on audio-visual features. In *Proc. of Int'l Conf. on Image and Video Retrieval*, volume 2383, pages 278–288, London, UK, July 2002. Springer, Lecture Notes in Computer Science.

[99] Z. Xiong, R. Radhakrishnan, A. Divakaram, and T. Huang. Audio events extraction based highlights extraction from baseball, golf and soccer games in a unified framework. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, Hong Kong, April 2003.

[100] Z. Xiong, R. Radhakrishnan, and A. Divakaram. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *IEEE International Conference on Image Processing (ICIP03)*, Barcelona,Spain, September 2003.

[101] W. Hua, M. Han, and Y. Gong. Baseball scene classification using multimedia features. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[102] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV formula 1 programs. In *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[103] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis using both audio and visual cues. *IEEE Signal Processing Magazine*, pages 12–36, November 2000.

[104] R. André-Obrecht, B. Jacob, and N. Parlangeau. Audio visual speech recognition and segmental master slave HMM. In *Audio-Visual Speech Processing Workshop*, Rhodes, Greece, September 1997.

[105] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled hmm for audio-visual speech recognition. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, 2002.

[106] P.S. Aleksic and A.K. Katsaggelos. Product hmms for audio-visual continuous speech recognition using facial animation parameters. In *IEEE International Conference on Multimedia and Expo (ICME03)*, Baltimore,MD, July 2003.

[107] J. Huang, Z. Liu, and Y. Wang. Integration of audio and visual information for content-based video segmentation. In *Proc. of IEEE International Conference on Image Processing (ICIP98)*, volume 3, pages 526–530, October 1998.

[108] J. Boreczky and L. Wilcox. A Hidden Markov Model framework for video segmentation using audio an image features. In *In Proc. of IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP98)*, Seattle, May 1998.

[109] H. Jiang, t. Lin, and H. Zhang. Video segmentation with the support of audio segmentation and classification. In *IEEE International Conference on Multimedia and Expo (I)(ICME00)*, volume 3, pages 1551–1554, August 2000.

[110] J-G. Kim, H. S. Chang, Y-T. Kim, K. Kang, M. Kim, J. Kim, and H-M. Kim. Multimodal approach for summarizing and indexing news video. *ETRI Journal*, 24(1) :1–11, February 2002.

[111] W. Qi, L. Gu, H. Jiang, X-R. Chen, and H-J. Zhang. Integrating visual, audio and text analysis for news video. In *Proc. of IEEE International Conference on Image Processing (ICIP00)*, Vancouver, September 2000.

[112] W. H-M. Hsu and S-F. Chang. A statistical framework for fusing mid-level perceptual features in news story segmentation. In *IEEE International Conference on Multimedia and Expo (I)(ICME03)*, volume 2, pages 413–416, July 2003.

[113] C. Saraceno and R. Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. In *Proc. of IEEE International Conference on Image Processing (ICIP98)*, pages 363–367, 1998.

[114] N. Adami, A. Bugatti, R. Leonardi, and P. Migliorati. Low-level processing of audio and video information for extracting the semantic content. In *IEEE Fourth Workshop on Multimedia Signal Processing*, pages 607–612, 2001.

[115] A. A. Alatan, A. N. Akansu, and W. Wolf. Multi-modal dialog scene detection using Hidden Markov Models for content-based multimedia indexing. *Multimedia Tools and Application*, 14(2) :137–151, 2001.

[116] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Communications of the ACM*, 40(12) :54–62, 1997.

[117] H. Sundaram and S.-F. Chang. Determining computable scenes in films and their structures using audio-visual memory models. In *ACM Multimedia*, pages 95–104, Los Angeles, California, November 2000.

[118] J. Huang, Z. Liu, and Y. Wang. Integration of multimodal features for video scene classification based on hmm. In *Proc. Of IEEE Workshop on Multimedia Signal Processing*, pages 53–58, Copenhagen, Denmark, September 1999.

[119] J. Huang, Z. Liu, and Y. Wang. Joint video scene segmentation and classification based on Hidden Markov Model. In *IEEE International Conference on Multimedia and Expo (I)(ICME00)*, volume 3, pages 1551–1554, August 2000.

[120] W.H. Adams, G. Iyengar, M.R. Naphade, C. Neti H.J. Nock, and J.R. Smtih. Semantic indexing of multimedia content using visual, audio and text cues. Technical report, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, November 2002.

[121] Y.-L. Chang, W.Zeng, I. Kamel, and R. Alonso. Integrated image and speech analysis for content-based video indexing. In *IEEE Conference on Multimedia Computing and Systems*, pages 306–313, 1996.

[122] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray. Automated generation of news content hierarchy by integrating audio, video, and text information. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP99)*, Los Alamitos, CA, USA, 1999.

[123] N. Babaguchi and N. Nitta. Intermodal collaboration : A strategy for semantic content analysis for broadcasted sports video. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[124] R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of sports program sequences by audio-visual analysis. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[125] Z. Liu and Q. Huang. Detecting news reporting using audio/visual information. In *Proc. of IEEE International Conference on Image Processing (ICIP99)*, volume 1, pages 324–328, October 1999.

[126] A. Hanjalic. Generic approach to highlights extraction from a sport video. In *IEEE International Conference on Image Processing (ICIP03)*, September 2003.

[127] M. Han, W. Hua, and Y. Gong. An integrated baseball digest system using maximum entropy method. In *Proc. ACM International Multimedia Conference*, pages 347–350, Juan-les-pins, France, December 2002.

[128] M. Betser, G. Gravier, R. Gribonval, and F. Bimbot. Extraction of information from video sound tracks - can we dectect simultaneous events ? In *Third International Workshop on Content-Based Multimedia Indexing (CBMI03)*, pages 71–77, September 2003.

[129] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual integration for tennis broadcast structuring. In *Third International Workshop on Content-Based Multimedia Indexing (CBMI03)*, pages 421–428, September 2003.

[130] S. Fine, Y. Singer, and N. Tishby. The Hierarchical Hidden Markov Model : Analysis and applications. *Machine Learning*, 32(1) :41–62, 1998.

[131] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Structuration multimodale d'une vidéo de tennis par Modèles de Markov Cachés. In *19e Colloque GRETSI sur le Traitement du Signal et des Images*, volume 3, pages 42–45, Paris, France, September 2003.

[132] E. Kijak, L. Oisel, and P. Gros. Hierarchical structure analysis of sport videos using hmms. In *IEEE Int. Conference on Image Processing (ICIP03)*, volume 2, pages 1025–1028, Barcelona, Spain, September 2003.

Résumé

Cette étude présente une méthode de structuration d'une vidéo utilisant des indices sonores et visuels. Cette méthode repose sur un modèle statistique de l'entrelacement temporel des plans de la vidéo. Le cadre général de la modélisation est celui des modèles de Markov cachés. Les indices visuels sont utilisés pour caractériser le type des plans. Les indices audio décrivent les événements sonores apparaissant durant un plan. La structure de la vidéo est représentée par un modèle de Markov caché hiérarchique, intégrant les informations a priori sur le contenu de la vidéo, ainsi que sur les règles d'édition. L'approche est validée dans le cadre des vidéos de tennis, ce dernier présentant une structure intrinsèque hiérarchique bien définie. En résultat de l'analyse de l'entrelacement temporel des différents types de plans, des scènes caractéristiques du tennis sont identifiées. De plus, chaque plan de la vidéo est assigné à un niveau de hiérarchie décrit en terme de point, jeu et set. Cette classification et segmentation simultanées de la structure globale de la vidéo peuvent être utilisées pour la création de résumés vidéo ou pour permettre une navigation non linéaire dans le document vidéo.

Mots-clefs : structuration vidéo, macro-segmentation, multimodalité, modèles de Markov cachés, modèles de Markov cachés hiérarchiques, analyse des vidéos de sport, indexation vidéo.

Abstract

This thesis is concerned with the structure analysis of sports videos using both audio and visual cues. The proposed method relies on a statistical model which takes into account both the shot content and the interleaving of shots. This stochastic modeling is performed in the global framework of Hidden Markov Models (HMMs) that can be efficiently applied to integrate prior information about video content and editing rules, and to merge audio and visual cues. Visual features are used to characterize the type of shot view. Audio features describe the audio events within a video shot. Our approach is validated in the particular domain of tennis videos, that present a hierarchical, complex and well-defined structure. The video structure parsing relies on the analysis of the temporal interleaving of video shots. Typical tennis scenes are simultaneously segmented and identified. In addition, each shot is assigned to a level in the hierarchy described in terms of point, game and set. As a result, the overall structure is identified. This can be used for video abstracting non-linear browsing of the document.

Keywords : video structure analysis, macro-segmentation, cross-modality, hidden Markov models, hierarchical hidden Markov models, sport video analysis, video indexing.